Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

AfriHuBERT: A self-supervised speech representation model for African languages

Jesujoba O. Alabi[1], *Xuechen Liu*[2], *Dietrich Klakow*[1], *Junichi Yamagishi*[2]

[1]Saarland University, Saarland Informatics Campus, Germany
[2]National Institute of Informatics, Japan

{jalabi,dietrich.klakow}@lsv.uni-saarland.de, {xuecliu,jyamagis}@nii.ac.jp

Abstract

In this work, we present AfriHuBERT, an extension of mHuBERT-147, a compact self-supervised learning (SSL) model pretrained on 147 languages. While mHuBERT-147 covered 16 African languages, we expand this to 1,226 through continued pretraining on 10K+ hours of speech data from diverse sources, benefiting an African population of over 600M. We evaluate AfriHuBERT on two key speech tasks, Spoken Language Identification (SLID) and Automatic Speech Recognition (ASR), using the FLEURS benchmark. Our results show a +3.6% F1 score improvement for SLID and a -2.1% average Word Error Rate (WER) reduction for ASR over mHuBERT-147, and demonstrates competitiveness with larger SSL models such as MMS and XEUS. Further analysis shows that ASR models trained on AfriHuBERT exhibit improved cross-corpus generalization and are competitive in extremely low-resource ASR scenarios.

Index Terms: Self-supervised learning, Multilingual speech representation, Speech processing, African languages

1. Introduction

Self-supervised learning (SSL)-based speech representation models such as HuBERT [1], XLS-R [2], and WavLabLM [3] have become an important component in the development of various speech-related applications, such as automatic speech recognition (ASR) [4, 5, 6], speech synthesis [7], speech translation [8], and spoken language understanding (SLU) [9]. These models, trained on vast amounts of unlabeled data, are designed to capture the nuances of different languages, enabling robust and accurate performance across diverse tasks.

Existing SSL models can be categorized as either monolingual or multilingual. Most monolingual SSL models are trained exclusively on English [1, 10], with only a few multilingual SSL models available, such as mHuBERT [11], w2v-XLSR [2], mHuBERT-147 [12], and MMS [13], which cover up to a thousand languages. While these models include some African languages, English and a few other high-resource languages often dominate the training data due to the abundance of available resources. Despite Africa's rich linguistic diversity, African languages remain relatively underrepresented. This lack of representation creates significant challenges in the building of robust speech-based dialog systems on the African continent, where thousands of languages and dialects coexist.

Some of the recent multilingual SSL models, such as MMS [13], w2v-BERT 2.0 [8], and XEUS [14], have demonstrated strong performance across different languages and tasks. However, these models tend to be large, with over 300 million parameters— making them computationally expensive and challenging to deploy in resource-constrained environments.

In contrast, mHuBERT-147, which is trained on 147 languages, including 16 African languages, offers a compact yet competitive alternative. Although it is competitive on benchmarks like ML-SUPERB [15], it lags behind on the FLEURS [16] benchmark for most languages, including most African languages. To address this gap, we introduce AfriHuBERT, the first massively multilingual and compact African-centric SSL model, built by extending mHuBERT-147 (95M parameters) through continued pretraining. AfriHuBERT is trained on a diverse dataset of 1,226 African languages and dialects, plus four widely spoken languages in Africa: Arabic, English, French, and Portuguese. The pretraining dataset is sourced from diverse domains, ensuring comprehensive phonetic and linguistic representation of the African context.

We evaluate our model on two downstream tasks, spoken language identification (SLID) and ASR, using FLEURS. AfriHuBERT significantly outperforms existing small SSL models on both tasks, including for languages with adaptation data solely from religious domains. This narrows the performance gap between mHuBERT-147 and large SSL models for African languages and, thereby highlighting the importance of tailored pretraining for speech representation in African languages. Our contributions are as follows.

1. We aggregate more than 10,000 hours of speech, covering more than 1,200 African languages and dialects.
2. We introduce AfriHuBERT, a multilingual SSL model for African languages, comparing continued pretraining and training from scratch over one iteration.
3. We evaluate AfriHuBERT against other multilingual SSL models on SLID and ASR, analyzing its predictions for both tasks.

We have released the pre-trained models of AfriHuBERT,[1][2] along with the codebase.[3]

2. Data and Pre-processing

2.1. Continued pretraining dataset

We aggregate data from various speech datasets across 1,230 languages, including 1,226 African languages and Arabic, English, French, and Portuguese. The four non-African languages are included to help preserve the model's ability on these languages and on their African-accented varieties. The data is gathered from 11 major sources: BibleTTS [17], Congolese Speech Radio Corpus (CSRC) [18], Jesus Dramas [14],

[1] https://huggingface.co/ajesujoba/AfriHuBERT
[2] https://doi.org/10.5281/zenodo.15531766
[3] https://github.com/nii-yamagishilab/AfriHuBERT

10.21437/Interspeech.2025-1437

Table 1: *Datasets used for training AfriHuBERT (after filtering and preprocessing the aggregated data), the amount of languages covered, the total duration, the domain of the data, the speech type, and licenses.*

Name	#Languages	Duration (h)	Domain	Type	License
BibleTTS [17]	6	357.6	Religious	Read	CC BY-SA 4.0
CSRC [18]	3	0.1	General	Radio	CC-BY
Jesus Dramas [14]	88	99.6	Religious	Read	CC BY-NC-SA 4.0
Kallaama [19]	3	124.9	Agriculture	Spontaneous	CC BY-SA 4.0
MCV [20]	4	1606.1	General	Read	CC-0
MMS ulab v2 [13, 14]	1230	2835.4	Religious	Read	CC BY-NC-SA 4.0
NaijaVoices [21]	3	1873.9	General	Read	CC BY-NC-SA 4.0
NCHLT [22, 23]	10	1889.4	General	Read	CC BY 3.0
Nicolingua [24]	10	142.4	News	Radio	CC BY-SA 4.0
VoxLingua107 [25]	13	886.4	General	Spontaneous	CC BY 4.0
Zambezi Voice [26]	5	176.0	General	Radio	CC BY-NC-ND 4.0

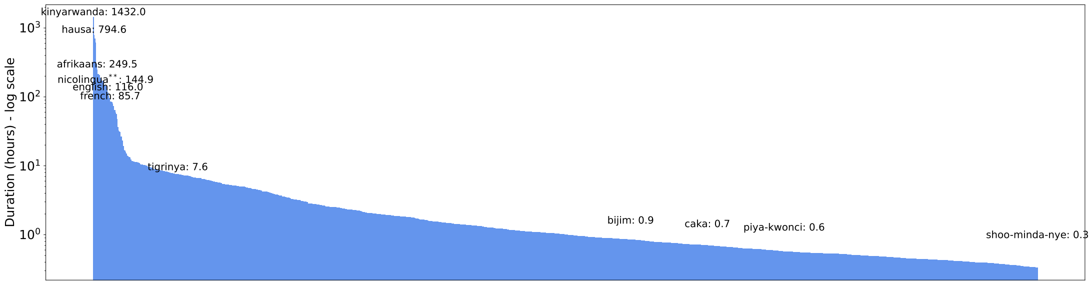

Figure 1: *Language-wise duration distribution of the aggregated data after preprocessing. Nicolingua** is not a language, but a collection of speech data for 11 languages, including ten from Guinea (a country in West Africa).*

Kallaama [19], Mozilla Common Voice (MCV) [20],[4] MMS ulab v2 [13, 14], NCHLT [22, 23], Nicolingua radio corpus [24], NaijaVoices [21] , VoxLingua107 [25], and Zambezi voice [26]. We filter Jesus Dramas and MMS ULAB v2 for African languages using GPT-4o[5] and Glottolog,[6] respectively; the former uses language names, the latter ISO 639-3 codes. Combining these data sources yields speech samples for 1,439 languages, including the four non-African languages. Table 1 summarizes the data sources and their properties.

2.2. Evaluation dataset

For downstream evaluation, we use the Sub-Saharan Africa (SSA) subset of the FLEURS dataset, which includes 20 languages from our pretraining data. We also include Kinyarwanda FLEURS,[7] and Arabic, English, French, and Portuguese, totaling 25 languages. We focus on SLID and multilingual ASR as downstream tasks.

2.3. Data preprocessing

All data including FLEURS were converted to single-channel audio and downsampled to 16 kHz. CSRC, Jesus Dramas, and MMS ulab v2 were segmented using WebRTC VAD.[8] The Kallaama dataset was split into manageable segments using transcription files provided by the authors. The noise in VoxLingua107 was filtered using the manifest from [12]. Additionally, we removed audio segments shorter than 1 second or longer than 30 seconds and excluded languages with less than 20 minutes of audio. As a result, the pretraining dataset contains over 10,000 hours of audio, covering 1,230 out of the 1,439 languages originally gathered. Figure 1 shows a skewed duration distribution, with Kinyarwanda accounting for over 10% of the total, while many languages have less than 10 hours of speech.

3. AfriHuBERT: Setup and Training

We train AfriHuBERT by extending mHuBERT-147 with the aggregated data, using multilingual adaptive finetuning (MAFT) [27, 28], a process of continued pretraining on multiple languages at once. Given the strong capabilities of mHuBERT-147, we use a one-iteration adaptation strategy [29]. Our objective is to answer two questions (1) *Can massively pre-trained mHuBERT-147 effectively generalize to African languages?* (2) *How effective is training AfriHuBERT from scratch using quality discrete targets from the pre-trained mHuBERT-147 without refinement?* Hence, we train three versions of AfriHuBERT. The first two require MAFT on mHuBERT-147 using its original discrete targets from the k-means model, while the other trains the k-means model on African language datasets to obtain AfriHuBERT-*o* and AfriHuBERT-*n*, respectively. Lastly, we train AfriHuBERT-*s* from scratch for one iteration using the new discrete targets. The new k-means model is trained using representations from the 9th layer of mHuBERT-147 with Faiss-based clustering [30]. We sampled up to 1 hour of speech data from each language and merged all samples to train the clustering model.[9]

To address language imbalance and ensure the model learns from underrepresented languages and dialects, we upsampled the aggregated data using temperature sampling with a multinomial distribution:

$$q_i = \frac{p_i^{\alpha}}{\sum_{j=1}^{D} p_j^{\alpha}}, \quad \text{where} \quad p_i = \frac{d_i}{\sum_{j=1}^{D} d_j}. \tag{1}$$

Here d is each language duration, D is the total number of languages, p is the probability of the language, and α is a tempera-

[4] version 17.0 hosted on Huggingface

[5] Version dated 2024-08-06.

[6] https://github.com/glottolog/glottolog

[7] https://huggingface.co/datasets/mbazaNLP/fleurs-kinyarwanda

[8] https://github.com/wiseman/py-webrtcvad

[9] Note that loading the entire dataset into memory is infeasible.

ture parameter that we set to 0.8.[10] We exclude English, Arabic, French, Portuguese, and audio data in Nicolingua.[11] Before upsampling, we allocate 10 minutes of audio for languages with less than 2 hours of data and 30 minutes for others as the validation set. We also include the original Nicolingua validation split. We train the models for 100K steps on upsampled data using the original HuBERT implementation within Fairseq [31]. Training uses a maximum of 128K tokens per batch, an update frequency of 64, and is optimized with a learning rate of 5e^{-5} and 32K warm-up steps.[12] The final models are selected based on the checkpoint with the lowest validation loss. Each model is trained using 4 NVIDIA A100 40GB GPUs.

4. Supervised Finetuning Setup

We evaluate AfriHuBERT models on SLID and multilingual ASR via supervised finetuning with FLEURS. Alongside mHuBERT-147, we evaluate a few other SSL models, including SSA-HuBERT [32] (95M params, trained on 20 African languages/dialects). We also fine-tune larger SSL models: XLSR-128 [2] and MMS [13] (316M), as well as w2v-BERT 2.0 [8] and XEUS [14] (580M). The latter two are not directly comparable due to their size and extensive pretraining; XEUS includes FLEURS in its training data, while w2v-BERT 2.0's pretraining data is undisclosed.

For SLID, we fine-tune on 25 FLEURS languages using an attentive static pooling layer, followed by a 512-D and 25-D softmax layer. Models are trained for 20 epochs with 3 random seeds; we report average F1 across all languages, both including and excluding the 4 non-African ones. To address imbalance, we cap each language at 1,030 samples, matching Afrikaans.

For multilingual ASR, models are fine-tuned jointly on all languages using CTC loss, without an external language model. The full fine-tuning setup uses a 3-layer FFN (1024 neurons, and LeakyReLU activation). We apply a 432-size character vocab from SentencePiece [33]. Training runs for 30 epochs with 3 random seeds, and we report the average overall WER. To ensure balanced training, we sample three hours of audio per language and merge the data across all 25 languages.

Following [12], we optimize fine-tuning for both tasks using Adam for the speech encoder, selecting the best learning rate from $1e^{-3}, 1e^{-4}, 1e^{-5}$. For SLID, the FFN uses Adam with a fixed learning rate of 0.001; for ASR, it uses Adadelta with a learning rate of 1.0, as implemented within SpeechBrain [34].[13] All models are trained on a single NVIDIA A100 GPU (40GB/80GB) with batch sizes of 32 (SLID) and 16 (ASR), using gradient accumulation as needed.

5. Results

Table 2 shows the SLID and ASR results, including average F1 and WER across all 25 languages and specifically for African languages. The following paragraphs summarize our key findings.

mHuBERT-147 is a strong, compact, multilingual SSL baseline. Overall, mHuBERT-147 outperforms SSA-HuBERT—a multilingual model of the same size—with a lower average WER, while SSA-HuBERT performs better in SLID. At the language level, we observed that SSA-HuBERT achieves better F1/WER on Hausa and Swahili, perhaps benefiting from pretraining on large datasets for both languages.

Table 2: *Performance of SSL models on FLEURS. We report the average F1 (%) and WER (%) scores for all languages (avg$_*$), and only the 21 African languages (avg). Size refers to each model's parameters (millions), and Dur denotes their pretraining data size (million hours).*

Models	**Size**	**Dur**	**SLID**(F1)↑		**ASR**(WER)↓	
	(M)	M(h)	avg$_*$	avg	avg$_*$	avg
Small SSL						
mHuBERT-147	95	$9e^{-2}$	88.0	85.8	50.4	52.1
SSA-HuBERT	95	$6e^{-2}$	89.6	88.0	56.6	56.2
AfriHuBERT-*s*	95	$1e^{-2}$	93.2	92.0	54.2	52.9
AfriHuBERT-*o*	95	$1e^{-2}$	90.3	88.9	48.4	49.3
AfriHuBERT-*n*	95	$1e^{-2}$	91.6	90.0	47.9	48.7
Large SSL						
w2v-XLSR	317	$4.4e^{-1}$	80.3	78.2	46.2	49.4
MMS	317	$4.9e^{-1}$	86.3	85.6	45.6	48.0
XEUS	577	$1.1e^{+1}$	**96.2**	**95.5**	46.5	49.5
w2v-BERT 2.0	580	$4.5e^{+1}$	92.7	91.3	**35.5**	**39.3**

Performing MAFT on mHuBERT-147 led to improved performance on African languages. On average, both AfriHuBERT-*o* and AfriHuBERT-*n* outperform the other two small SSL models on both tasks, while achieving comparable performance. Compared to mHuBERT-147, both models perform better on all African languages except Arabic, English, French, and Portuguese, which were dominant during pretraining but underrepresented during adaptation. Languages like Luo and Kimbundu, which were not present during pretraining and introduced only during adaptation with a few hours of religious data, show improvements over mHuBERT-147. Also, AfriHuBERT-*s* outperforms all small models on SLID but is not competitive to mHuBERT-147 or other AfriHuBERTs for ASR. We hypothesize that training AfriHuBERT-*s* for longer steps might improve its representation to the point that it can match mHuBERT-147's performance on African languages.

w2v-BERT 2.0 is a large competitive model. Among large SSL models, w2v-BERT 2.0, trained on more than 4.5M hours of audio, achieves the best overall performance on ASR due to its size and data volume, while XEUS delivers the best SLID performance but significantly lags behind w2v-BERT 2.0 in ASR. MMS and w2v-XLSR, with similar parameter counts and pretraining data, perform competitively, with MMS showing a slight improvement.

6. Analysis & Discussion

Going forward, we focus on AfriHuBERT-*n*, now called AfriHuBERT. We analyze its failure cases on both tasks, assess cross-corpus ASR generalization, and evaluate its performance in extremely low-resource and multi-dialect ASR, comparing it to mHuBERT-147 and MMS.

SLID confusion matrix for AfriHuBERT: Inspecting AfriHuBERT's confusion matrix reveals that geographically close languages are often misclassified as each other. For example, 40% of the audio samples speaking Zulu, a South African language, are misclassified as Xhosa on average. However, this miss-classification does not occur in the reverse direction. We hypothesize that this stems from training data artifacts or linguistic similarities, which future work can explore. Similarly, Fulfude, spoken in West and Central Africa, is misclassified as

[10] It is computationally expensive to test all possible values of α.

[11] These exclusions were due to the inability to separate the audio data into the ten respective languages including French.

[12] lr = 5e^{-3} when training from scratch.

[13] We adapted the IEMOCAP and VoxLingua107 SpeechBrain recipes for SLID, and the DVoice recipe for ASR.

Hausa, Somali, or Wolof, which are languages from overlapping regions.

Error analysis of the multilingual ASR output: Based on the ASR results, we analyze system outputs, with a focus on AfriHuBERT. Using Yorùbá (the language with the second-highest WER) as a case study, we identify data quality issues, likely due to error propagation from the FLORES-101 [35] dataset, the source of FLEURS. A manual inspection of the Yorùbá transcriptions shows that they did not follow the standard Yorùbá orthography with instances without diacritics, or a mixture of diacritics and no diacritics. For example:

(1) **Groundtruth transcription:** won se ikede naa leyin ti trumpi ba aare toki resep tayipi edogani lori ago
When diacritized: wọ́n ṣe ìkéde náà lẹ́yìn tí trumpi bá ààrẹ toki resep tayipi edogani lórí ago
Translation: they made the announcement after trump had president toki resep tayipi edogani on a phone call
AfriHuBERT: wọ́n se ìkéde náà lẹyìn tí tromp b arẹ toki recept tayipà èdògáni lórí ago

Example 1 shows a ground-truth transcription that does not follow Yorùbá orthography and is completely undiacritized, while AfriHuBERT's output is partially diacritized. These inconsistencies, especially in the FLEURS training data, likely contribute to the Yorùbá ASR models' high WER. Beyond diacritics, the models also struggle with transcribing named entities. Future work should further audit FLEURS transcriptions and correct these errors, similar to FLEURS-R [36], which focused on improving FLEURS's audio quality.

Table 3: *Cross-corpus generalization of ASR models on MCV.*

	afr	amh	hau	ibo	kin	lug	swh	yor	Avg
CER (%)									
mHuBERT-147	15.2	46.2	17.4	21.1	24.6	18.1	19.6	38.8	25.1
AfriHuBERT	13.2	**42.5**	**14.1**	18.3	**22.3**	**16.6**	17.6	**35.8**	**22.6**
MMS	**13.1**	48.7	16.3	**17.0**	24.4	17.2	**17.3**	37.2	23.9
WER (%)									
mHuBERT-147	53.1	85.8	59.4	62.3	72.4	71.3	59.0	86.9	68.8
AfriHuBERT	48.0	**81.0**	**51.1**	60.5	**66.5**	**67.4**	**52.6**	**81.2**	**63.6**
MMS	**43.6**	83.7	57.9	**56.2**	72.1	70.6	53.0	84.4	65.2

Cross-corpus ASR generalization of AfriHuBERT: Next, we evaluate the ASR models from AfriHuBERT, mHuBERT-147, and MMS—three multilingual HuBERT-style models—on another ASR corpus to assess their cross-corpus generalization. For this, we used the MCV [20] test split,[14] which covers eight of the 21 African languages they were originally trained on. The CER and WER results presented in Table 3 show that AfriHuBERT, on average, generalizes better out-of-domain and outperforms mHuBERT-147 and MMS with WERs of 68.8% and 65.2%, respectively. Specifically, AfriHuBERT achieves a WER of less than 60% in three languages (Afrikaans, Hausa, Swahili) and of less than 70% for Igbo, Kinyarwanda and Luganda. In contrast, Amharic (using a non-Latin script) and Yorùbá (using Latin script with diacritics) have WERs of 81.0% and 81.2%, respectively.

Evaluating AfriHuBERT on low-resource multilingual ASR: Furthermore, we evaluate the three SSL models in extremely low-resource settings. Using the experimental setup from Section 3, we fine-tune the models for multilingual ASR with 10 and 30 minutes of audio data per language, respectively, and evaluate their performance on African languages only.

The results in Figure 2 show that, on average, AfriHuBERT outperforms mHuBERT-147 in both settings and remains competitive with MMS. In the 10-minute setting, AfriHuBERT achieved a WER of 65.7% and a CER of 20.5%, compared to mHuBERT-147 's 73.0% and 23.1%, respectively. In the 30-minute setting, WERs considerably decreased across all three models. However, for most languages, WERs remained above 60%, with only a few exceptions, highlighting the need for more data to improve ASR performance for these languages.

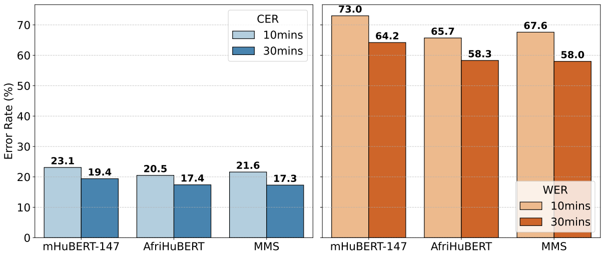

Figure 2: *ASR performance in extremely low-resource ASR scenarios.*

Table 4: *Multi-dialect ASR performance comparison on YORÙLECT (comparing three Yorùbá dialects).*

Models	**Standard**		**Ife**		**Ilaje**		**Avg**	
	CER	WER	CER	WER	CER	WER	CER	WER
mHuBERT-147	11.9	40.8	22.4	65.1	17.1	51.0	17.1	52.3
AfriHuBERT	**11.2**	**37.7**	**21.4**	62.9	16.4	48.8	**16.3**	49.8
MMS	11.4	38.2	21.6	**62.5**	**15.8**	**47.5**	**16.3**	**49.4**

Multi-dialect Yorùbá ASR evaluation: Lastly, given the flaws identified in Yorùbá FLEURS, we address whether our findings, particularly the observed improvements and competitiveness by AfriHuBERT, can be trusted. We train a multi-dialect Yorùbá ASR model on the well-curated YORÙLECT [37] dataset using a similar setup as before but with a sentencepiece character vocabulary of size 63. We focus on three dialects: Standard Yorùbá, Ife, and Ilaje. Our results in Table 4 show that, on average and across dialects, AfriHuBERT outperforms mHuBERT-147, and is competitive to MMS. We hypothesize that models perform best on the standard dialect due to its abundant resources, while Ife is the most challenging due to its scarcity. These results confirm that, despite AfriHuBERT's compact size, it is still competitive.

7. Conclusion

In this work, we created AfriHuBERT by extending mHuBERT-147 to 1,226 African languages via MAFT on speech data aggregated from various sources, including these languages and four widely spoken non-indigenous languages in Africa. We evaluated both compact SSL models such as mHuBERT-147 and AfriHuBERT together with some other large multilingual SSL models on both SLID and ASR tasks and found that mHuBERT-147 is a strong multilingual SSL baseline. However, AfriHuBERT which is an extension of mHuBERT-147, outperforms other SSL models on average for both tasks. In future work, we plan to upscale AfriHuBERT, enhancing its generalizability to better accommodate unseen African languages and dialects.

8. Acknowledgements

This work was conducted during the first author's internship at NII, Japan. This study is partially supported by JST AIP Acceleration Research (JPMJCR24U3). Part of this study was carried out using the TSUBAME4.0 supercomputer at the Institute

[14] version 17.0

of Science Tokyo. Also, we thank Xin Wang, Badr M. Abdullah, Siyang Wang, Wanying Ge, David Adelani, and Aravind Krishnan for their helpful feedback.

9. References

[1] W.-N. Hsu, B. Bolte *et al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[2] A. Babu, C. Wang *et al.*, “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” *arXiv*, vol. abs/2111.09296, 2021.

[3] W. Chen, J. Shi *et al.*, “Joint prediction and denoising for large-scale multilingual self-supervised learning,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[4] S. Pascual, M. Ravanelli *et al.*, “Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks,” in *Proc.INTERSPEECH*, 2019, pp. 161–165.

[5] Y.-A. Chung, Y. Zhang *et al.*, “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)*. IEEE, 2021, pp. 244–250.

[6] Y. Zhang, W. Han *et al.*, “Google usm: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.

[7] C. Gong, X. Wang *et al.*, “Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 4036–4051, Sep. 2024.

[8] S. Communication, L. Barrault *et al.*, “Seamless: Multilingual expressive and streaming speech translation,” 2023. [Online]. Available: https://arxiv.org/abs/2312.05187

[9] Y. Peng, S. Arora *et al.*, “A study on the integration of pre-trained ssl, asr, lm and slu models for spoken language understanding,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 406–413.

[10] A. Baevski, Y. Zhou *et al.*, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] A. Lee, H. Gong *et al.*, “Textless speech-to-speech translation on real data,” *arXiv preprint arXiv:2112.08352*, 2021.

[12] M. Zanon Boito, V. Iyer *et al.*, “mhubert-147: A compact multilingual hubert model,” in *Interspeech 2024*, 2024, pp. 3939–3943.

[13] V. Pratap, A. Tjandra *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[14] W. Chen, W. Zhang *et al.*, “Towards robust speech representation learning for thousands of languages,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 10 205–10 224.

[15] J. Shi, D. Berrebbi *et al.*, “Ml-superb: Multilingual speech universal performance benchmark,” in *Interspeech*, 2023.

[16] A. Conneau, M. Ma *et al.*, “Fleurs: Few-shot learning evaluation of universal representations of speech,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[17] J. Meyer, D. Adelani *et al.*, “Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus,” in *Interspeech*. ISCA, 2022.

[18] U. Kimanuka, C. wa Maina *et al.*, “Speech recognition datasets for low-resource congolese languages,” *Data in Brief*, vol. 52, p. 109796, 2024.

[19] E. Gauthier, A. Ndiaye *et al.*, “Kallaama: A transcribed speech dataset about agriculture in the three most widely spoken languages in senegal,” in *Proceedings of the Fifth workshop on Resources for African Indigenous Languages (RAIL 2024)*, 2024.

[20] R. Ardila, M. Branson *et al.*, “Common voice: A massively-multilingual speech corpus,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[21] C. Emezue, T. N. Community *et al.*, “The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages,” 2025. [Online]. Available: https://arxiv.org/abs/2505.20564

[22] E. Barnard, M. H. Davel *et al.*, “The nchlt speech corpus of the south african languages,” in *Workshop Spoken Language Technologies for Under-resourced Languages (SLTU)*, 2014.

[23] J. Badenhorst and F. De Wet, “Nchlt auxiliary speech data for asr technology development in south africa,” *Data in Brief*, vol. 41, p. 107860, 2022.

[24] M. Doumbouya, L. Einstein *et al.*, “Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021.

[25] J. Valk and T. Alumäe, “Voxlingua107: a dataset for spoken language recognition,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[26] C. Sikasote, K. Siaminwe *et al.*, “Zambezi Voice: A multilingual speech corpus for zambian languages,” in *Proc. INTERSPEECH 2023*, 2023, pp. 3984–3988.

[27] Y. Tang, C. Tran *et al.*, “Multilingual translation with extensible multilingual pretraining and finetuning,” *arXiv preprint arXiv:2008.00401*, 2020.

[28] J. O. Alabi, D. I. Adelani *et al.*, “Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning,” in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4336–4349.

[29] J. Xu, M. Wu *et al.*, “Seamless language expansion: Enhancing multilingual mastery in self-supervised models,” *arXiv preprint arXiv:2406.14092*, 2024.

[30] M. Douze, A. Guzhva *et al.*, “The faiss library,” 2024.

[31] M. Ott, S. Edunov *et al.*, “fairseq: A fast, extensible toolkit for sequence modeling,” in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[32] A. Caubrière and E. Gauthier, “Africa-centric self-supervised pre-training for multilingual speech representation in a sub-saharan context,” in *5th Workshop on African Natural Language Processing*, 2024.

[33] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[34] M. Ravanelli, T. Parcollet *et al.*, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.

[35] N. Goyal, C. Gao *et al.*, “The Flores-101 evaluation benchmark for low-resource and multilingual machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022.

[36] M. Ma, Y. Koizumi *et al.*, “Fleurs-r: A restored multilingual speech corpus for generation tasks,” in *Interspeech 2024*, 2024, pp. 1835–1839.

[37] O. Ahia, A. Aremu *et al.*, “Voices unheard: NLP resources and models for Yorùbá regional dialects,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 4392–4409.