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

Motivation

- Self-supervised learning (SSL) speech representation models are important component of various speech-related systems
- African languages remain relatively underrepresented in existing SSL models
- Multilingual SSL models like w2v-BERT 2.0, and XEUS perform well across languages and tasks but are large
- Can we build a **compact** SSL model for African languages?🤔

Research questions

1. Can massively pretrained mHuBERT-147 effectively generalize to African languages (to have AfriHuBERT)?
2. Can pre-training from scratch be effective using mHuBERT-147 targets without refinement?

Dataset for Pre-training

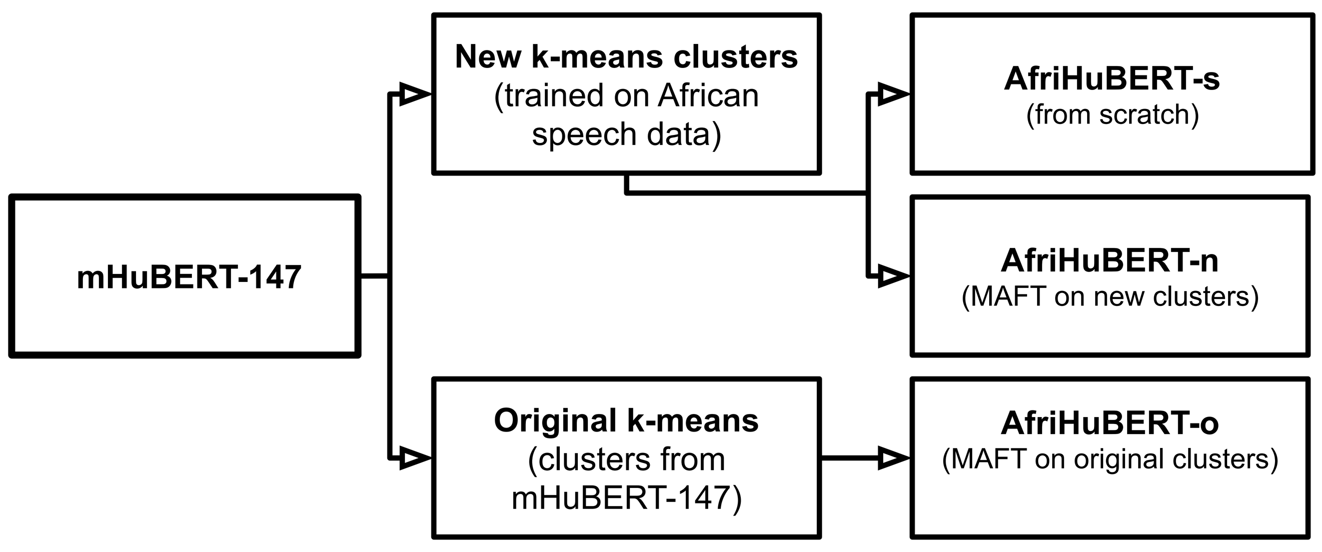
- We aggregate data from 11 major sources
- Combining these data sources return audio samples for 1,435 languages
- We exclude languages with less than 20m of audio, to have 1,226 African languages
- We include Arabic, English, French, and Portuguese from MMS ulab v2 only
- Only about 64 languages had more than 10 hours of audio samples

Name	#Langs.	Dur. (h)	Domain	Type	License
BibleTTS	6	357.6	Religious	Read	CC BY-SA 4.0
CSRC	3	0.1	General	Radio	CC-BY
Jesus Dramas	88	99.6	Religious	Read	CC BY-NC-SA 4.0
Kallaama	3	124.9	Agriculture	Spontaneous	CC BY-SA 4.0
MCV	4	1606.1	General	Read	CC-0
MMS ulab v2	1230	2835.4	Religious	Read	CC BY-NC-SA 4.0
NaijaVoices	3	1873.9	General	Read	CC BY-NC-SA 4.0
NCHLT	10	1889.4	General	Read	CC BY 3.0
Nicolingua	10	142.4	News	Radio	CC BY-SA 4.0
VoxLingua107	13	886.4	General	Spontaneous	CC BY 4.0
Zambezi Voice	5	176.0	General	Radio	CC BY-NC-ND 4.0

Datasets used for training AfriHuBERT.

AfriHuBERT: Pretraining Setup

- We train 3 variants of AfriHuBERT



AfriHuBERT variants: S (from-scratch), N (MAFT with new clusters), O (MAFT with original clusters).

AfriHuBERT: Training and Evaluation

- Adaptation and training done for 100K steps
- All models were fine-tuned and evaluated on Spoken LID and ASR using the Sub-Saharan African subset of FLEURS
- We also include English, Arabic, French and Portuguese
- Pretraining with Fairseq, evaluation with SpeechBrain
- As baseline, we compare to both small (African-centric) and large SSL models
- ⚠️ Evaluation covers only 2% of training languages

TL;DR

1. We aggregate 10K hours of speech from 1,200+ African languages to build a compact SSL model, AfriHuBERT
2. AfriHuBERT benefits from mHuBERT-147's multilingual foundation and multilingual religious data
3. AfriHuBERT outperforms similar-sized SSL models on speech tasks and competes with larger ones
4. FLEURS transcriptions require auditing and corrections

Results: SLID & ASR

Models	Size (M)	Dur M(h)	SLID(F1)↑ avg*	SLID(F1)↑ avg	ASR(WER)↓ avg*	ASR(WER)↓ 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**

Performance of the SSL models on FLEURS. We report the average F1 (%) and WER (%) scores for all languages (avg*), and 21 African languages (avg).

Findings

- mHuBERT-147 is a strong, compact, multilingual SSL baseline
- MAFT on mHuBERT-147 using primarily religious speech improved performance on all 21 African languages
- New pseudo-labels → Slight gain in AfriHuBERT performance over the original

Error Analysis of SLID outputs

- We inspected the AfriHuBERT's SLID confusion matrix
- Geographically close languages (e.g., Xhosa–Zulu, Fulfulde–Wolof–Hausa) are often misclassified as each other

Error Analysis of ASR outputs

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

FLEURS groundtruth transcriptions are inaccurate for Yoruba. 🤔

Can we trust our results? Yes! ☺️

Models	Standard CER	Standard WER	Ife CER	Ife WER	Ilaje CER	Ilaje WER	Avg CER	Avg 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 ASR performance comparison on YORÙLECT (comparing 3 Yorùbá dialects.).