VoxMg: An Automatic Speech Recognition Dataset for Malagasy

Falia Ramanantsoa

Mathématiques, Informatique et Statistique Appliquées (MISA), University of Antananarivo

African languages are not well-represented in Natural Language Processing (NLP). The main reason is a lack of resources for training models. Low-resource languages, such as Malagasy, cannot benefit from modern NLP methods if no datasets are available. This paper presents the curation and annotation of VoxMg, a speech dataset for Malagasy that consists of 3873 audio files totaling 10.80 hours. We also run a baseline, which is the first Automatic Speech Recognition (ASR) model ever built in this language and obtained a Word Error Rate (WER) of 33%.

Motivations

- ASR has become more ubiquitous in various aspects of our daily lives such as voice assistants and automatic captioning.
- ASR has made significant progress over the past years, especially for high resource languages like English.
- Many efforts have been made for low-resource languages, such as African languages, but ***Malagasy has not been included yet***.
- ***There is no ready-to-use data for Malagasy***.

Creation of VoxMg Dataset

VoxMg is the first ever ready-to-use speech dataset for Malagasy, lasting **10.80 hours** with **3873 audio files** and created using two sources: **VoxLingua107** and **nybaiboly.net**:

- 57% of the 9-second and 87% of the 10-second audio files from VoxLingua107, giving **3475 files** to transcribe and totaling **9.24 hours**.
- To prevent model overfitting on the single speaker from nybaiboly.net, 30 chapters from Genesis giving **820 files** to transcribe and lasting **2.20 hours**.

Statistics of VoxMg Dataset

Corpus	VoxLingua107	nybaiboly.net	Total
Treated (hrs)	9.24	2.58	11.82
Validated (hrs)	8.60	2.20	10.80
Audio Chunks	3053	820	3873
Audio Chunks per Gender (M—F)	2229—824	820—0	3049—824
Num. Speakers (M—F)	306—153	1—0	307—153

Subset	Train	Validation	Test
Duration (hrs)	8.63	1.09	1.08
Audio Chunks per Gender (M—F)	2446—655	299—87	304—82
Num. Speakers (M—F)	113—85	86—55	195—57

WER on Wav2Vec2-XLS-R-300M

Model	WER
Wav2Vec2-XLS-R	44.39
Wav2Vec2-XLS-R + LM	33.43

Discussion and Future Works

- Importance of Language Model (LM)
- Spontaneous speech
- Increase of quantity and diversity
- Including dialects

References

[1] 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. [2] “Malagasy Biblical Media,” retrieved : 2022-11-08. [Online]. Available : https://nybaiboly.net/. [3] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised crosslingual speech representation learning at scale. arXiv preprint arXiv:2111.09296, 2021.