Language Info

	Language compiled	Region	ISO Code	Who Collected the Data	Dialect	Dataset size
	Wolof	West Africa(Senegal)	wo	Amina Rufai &Mossou Morry Fadillou	Baol	464 Utterances

Motivation

As graduate students at African Institute of Mathematical Sciences/Machine Intelligence(AIMS/AMMI), part of the experiences encountered outside academics is the difficulty of communication amongst native dialect speakers and students from other parts of Africa who speak a different language. Communication centered around Basic conversations, navigating across cities(boarding a taxi), eating at a restaurant, going to the doctor/a visit to the hospital, et cetera, is a major challenge, without having a bilingual speaker(who speaks both the languages). This is the motivation for this research. Also, Wolof is a major language not just in Senegal but also in other neighbouring countries such as, The Gambia, and Mauritania, is the native language of the ethnic group of the Wolof people, and it is a Low resource language. It is estimated that about 3.2 million people speak Wolof as a first language and approximately 3.5 million second language speakers.

Composition

A total of records using a data dictionary built from scratch. The data dictionary was built with conversations around these phrases: Basic conversations, navigating across cities(boarding a taxi), eating at a restaurant, going to the doctor/a visit to the hospital.

Gender: Female
Age Range: 30-35

Collection process (where/how)
Where and How: Text documentation on collaborative google docs, and via audio recorder.

Source of prompts: Audio Voice recorded downloaded from Google playstore

Recommended uses/applications

While building this data, one of our objectives was to build a conversational AI system that can translate between english and wolof, while also serving as language guide for tourists, students visiting Senegal for the first time.

Recommended Applications:

- Speech-to-text and a translation system working like Google translate
- Applied to a speaker diarization system for wolof
- Keyword Extraction system
- Autocomplete, Voice pronunciation mobile keyboards
- Contribution of dataset which would be useful to the community

Model Information

Model Used	Fine-tuned on Facebook Wav2vec: “facebook/wav2vec2-large-xlsr-53”
Total Data Size	464
Total number of Audio hours	0.2 hours / 7 minutes approx (average of 2seconds long per audio)
Training Data Size	298

Val Data size	74
Test Data Size	92
Learning Rate	1e-4
Num_Epochs	100

Model results

Wer: 0.9
Training loss:3.674060299521998
Validation loss: 1.411783

Issues Captured:

- More data compilation required to train model
- Proper data formatting and cleaning