Design and Evaluation of a Yoruba Dialogue Transcription System

Alberto Villalvazo, Jonathan Mukiibi, Lingyun Gao, Lu Han, Oluwadunsin

Mentor: Thomas Thebaud

Problem Statement

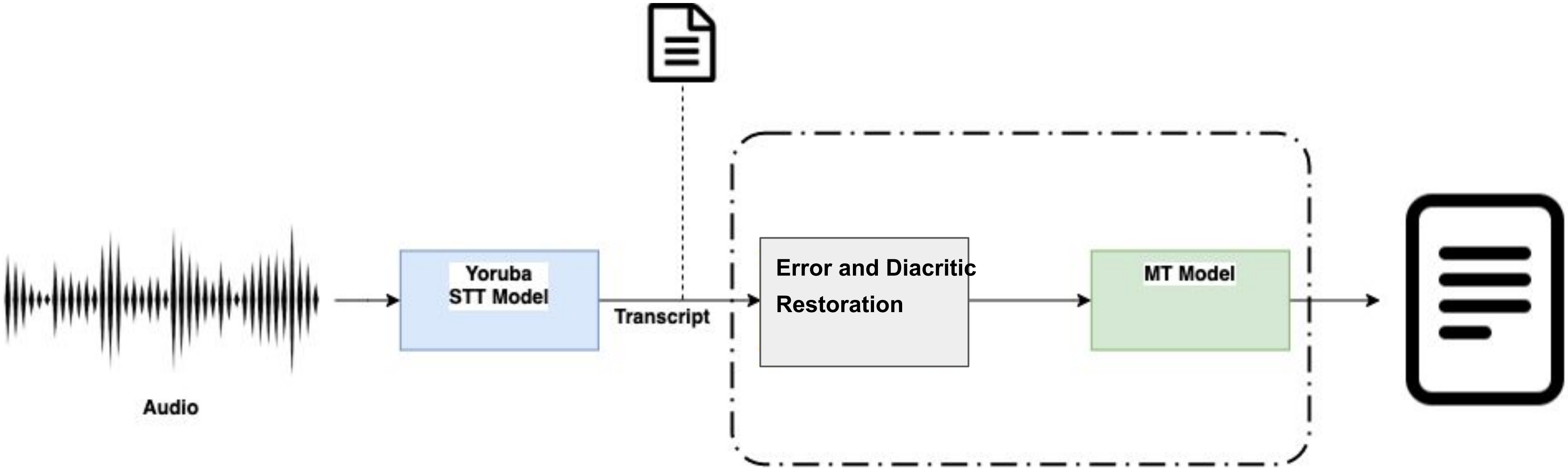
- Yoruba: 3rd most spoken language in Africa.
- Spoken by over 40 million native speakers
- Typical low-resource language:
  - Lacking labeled corpus,
  - Lacking applications like auto speech recognition and machine translation

- The goal: develop a Yoruba Dialogue Transcription System
  - that allows communication with computer systems and non-speakers.

Solution:

- Develop a Yoruba Dialogue Translation System including three parts:
  - Speech to Text Module (xlsr-53 wav2vec)
  - Automatic Damage Restoration (MarianMT)
  - Machine Translation (MarianMT)
- Leverage 4 Yoruba datasets
- Finetune multilingual transformer model

Solution : Yoruba Dialogue System



Results:

Model	Result	State-of-art
STT	WER:51.7	WER 45.6
ADR(error dataset)	WER 52.0->26.2	----
MT	BLEU metric 42.8	BLEU metric 22.4±0.5

Future Work

- End2End Training：Intergrate three parts to improve global performance
- Data collection: We are lacking data for evaluating an end2end system