Bias against English-Speaking Africans in Automated Speech Recognition

William Lu and Aditya Singh

Motivation

- ➔ Automated Speech Recognition (ASR) systems are widely used in various applications such as virtual assistants (e.g., Siri, Alexa), voice commands on mobile devices, and automated job interview processes.
- ➔ Research indicates that ASR technology tends to be less effective for African Americans and other English-speaking African groups, demonstrating a higher error rate in recognizing their speech compared to white speakers.
- ➔ Addressing these biases is crucial, underscoring the need for increased research efforts to make ASR systems more equitable and inclusive.
- ➔ Improve the accuracy of ASR systems for African English speakers to ensure equitable technology access.

Research Question

➔ How can the performance of Automated Speech Recognition (ASR) systems be improved for African English speakers through the fine-tuning of OpenAI's Whisper ASR system?

Approach

- **Fine tuning Whisper Small ASR model with AfriSpeech-200 dataset (South African English + English accent)**
- Why the use of AfriSpeech-200 data?
  - To compensate the lack of representation of English spoken by African
- Preprocessing
  - A Whisper Feature Extractor
    - Audio samples were standardized to a sampling rate of 16kHz, consistent with Whisper-small, to maintain correct audio speed.
    - Audio samples were either padded or truncated to ensure each has a consistent length of 30 seconds
    - Converted the adjusted audio samples into log-Mel spectrograms, which better mimic the human auditory range, facilitating more effective model training.
- Model Building
  - Whisper-small uses a sequence-to-sequence architecture, converting audio spectrogram features into sequences of tokens
    - It utilizes a Whisper feature extractor to transform input audios into log-Mel format.
    - Encoder blocks process these log-Mel audios into hidden states. Decoders predict tokens and use cross-attention complete transcriptions.
    - A data collector was implemented to transform preprocessed data into PyTorch tensors for model training
- Evaluation Metric
  - Word Error Rate (WER): # correct words/# total words

Results

Model / Data	With Fine Tuning	Without Fine Tuning
South African English Accent Audios	23.65	25.53
English Accent Audios	53.28	49.13

Conclusion + Future Steps

- **Conclusion**
  - Fine tuning with more underrepresented data can improve ASR performance overall
  - Data quality matters

- **Future Steps**
  - Fine tuning Whisper-small with more data
  - Train the ASR system with more representative data at first

Thank You!