SABC2TXT

Creating high-quality text corpora for low-resource languages

Overview

IsiXhosa is considered a low-resource language because of its limited availability of electronic documents. Natural language processing techniques require electronic documents to extract content for processing to enable computers to understand, interpret and generate spoken languages. Collecting and producing these materials is often time-consuming, costly, and often not feasible for low-resource languages. This project aims to assess the viability of automatically transcribing audio to create a high-quality textual corpus using standard speech tools and resources obtained from SADiLaR.

Gold Standard ASR Corpus

Research question: Can crowdsourcing via a web application effectively produce a high-quality gold standard Automatic Speech Recognition corpus for isiXhosa from audio data collected from casual conversations and broadcast news?

Results: Two sub-corpora were developed for two audio sets—structured and unstructured. Statistical analysis on both corpora revealed:

- An above-average **inter-transcriber similarity** score of approximately 60%.
- A **token-to-type ratio** of approximately 2.5.
- Adherence to **Zipf's law** as seen in natural languages.

Conclusions: Despite limitations like small sample sizes, the corpus is a notable effort in addressing the shortage of electronic linguistic resources in African languages.

Dataset Statistics

Description	Structured Corpus	Unstructured Corpus
Audio Segments	84	45
Transcriptions	45	22
Total Words	26794	12764
Distinct Words	9837	5165
Most Appeared Word	Ukuba	Uba
Least Appeared Word	Minist	Iyandi

SABC News Audio Transcription

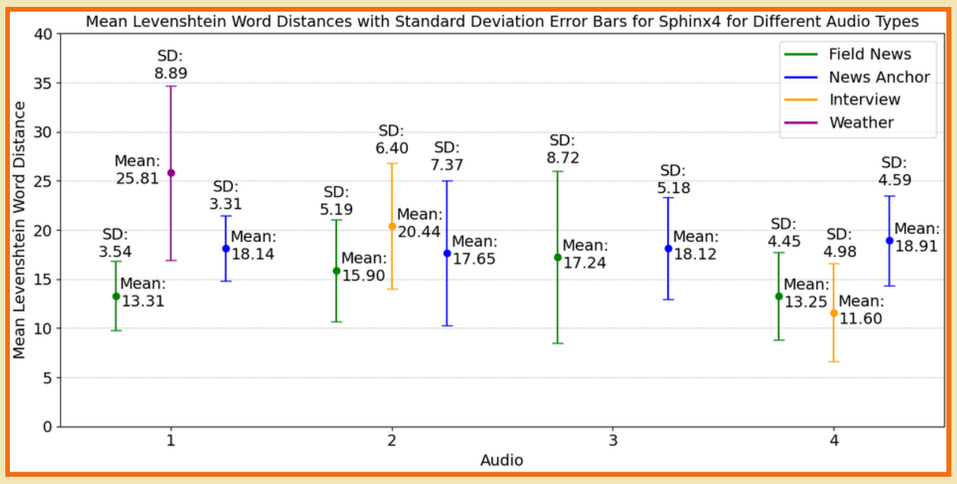
Research question: How accurate is the use of the CMUSphinx speech recognition toolkit (Sphinx4 and Pocketsphinx) for transcribing isiXhosa SABC news?

Results:

- Pocketsphinx performed better than Sphinx4.
- Recognition for male speakers was better than female speakers.
- Recognition for field news anchor was better than the main news anchor (shown in the figure on the right).

Conclusions: Overall CMUSphinx performed poorly on unseen data, therefore other speech recognition toolkits should be explored as well as increasing the amount of data used to train models.

Different Types of Speech in 4 Audios



Mobile Audio Transcription

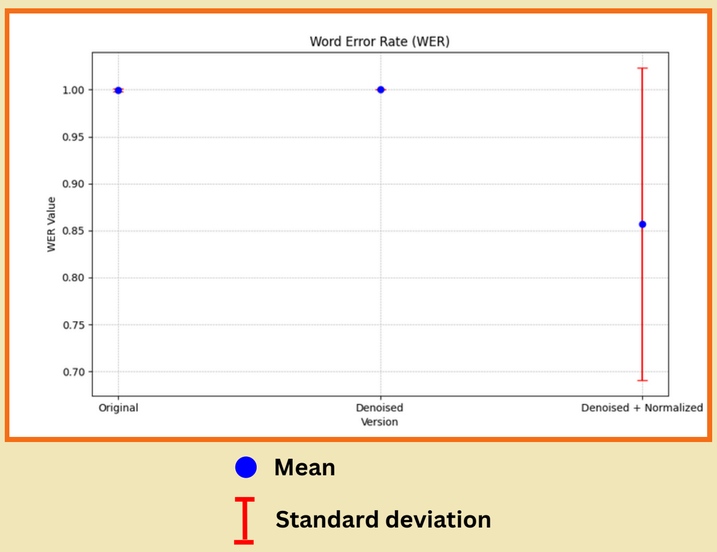
Research question: How accurate is speech-to-text transcription of unstructured audio using the PocketSphinx speech recognition toolkit on mobile devices?

Results:

- The system produced high Word Error Rates and Levenshtein distances for casual speech due to factors such as code-switching and background noise.
- Normalization and reducing the volume (dB) of background noise in audio improved transcription accuracy (shown in the figure on the right).
- Hardware differences such as microphone quality and the rate of speech also influence performance.

Conclusions: The tool demonstrated a poor performance on unstructured audio, highlighting a need for specialized model improvements.

Casual speech



Project team:

Sabiha Shaikh (shkmos004@uct.ac.za)
Kristen Jodie Basson (bsskri003@uct.ac.za)
Fardoza Tohab (thbfar002@uct.ac.za)

Supervisors:

Prof. Hussein Suleman
Dr. Abayomi Agbeyangi

University of Cape Town
Department of Computer Science
Email: dept@cs.uct.ac.za
Tel: 021 650 2663