Automatic Speech Recognition for Nigerian-Accented English

Oreoluwa Boluwatife Babatunde [1] Emmanuel O. Akeweje [2] Sharon Ibejih
Victor Tolulope Olufemi [3] Sakinat Oluwabukonla Folorunso [1]

[1]Olabisi Onabanjo University [2]Trinity College Dublin [3]Obafemi Awolowo University

babatundeoreoluwa35@gmail.com

Abstract

This research aims to address the gap in the performance of ASR systems on low-resourced English accents by using publicly available Nigerian-accented English data. By creating ASR models capable of accurately interpreting and transcribing Nigerian-accented and contextual English, we strive to ensure equitable access to ASR technologies and services for English speakers with Nigerian accents. The experiment in this study employs transfer learning techniques on NeMo's QuartzNet15x5 English model and Wav2vec2.0 XLS-R300M. The best model generated a word error rate 8.2%, outperforming the free Google Speech Recognition library of 44.2% WER on the Nigerian English accent test data.

Introduction

Recognizing the importance of fair and accurate access to ASR technology for individuals with African accents, there has been a growing interest in developing ASR systems specifically designed to distinguish and transcribe African-accented speech. The focus is on creating robust ASR models capable of accurately transcribing and interpreting speech across a wide range of accents and dialects. The goal of this project is to develop an end-to-end ASR system specifically tailored to Nigerian-accented English. By addressing the unique challenges and characteristics of Nigerian accents, the aim is to ensure equitable access to ASR technologies and services for individuals with Nigerian accents. This project seeks to contribute to the development of ASR models that understand and transcribe Nigerian-accented English accurately, fostering inclusivity and enhancing communication for diverse linguistic communities

Methodology

• **Data Collection**: The dataset used in this research is a combination of openly accessible Google Nigerian speech data [6] and SautiDB's Nigerian English data [7], explained in further detail below. Both datasets gave a typical representation of how the average Nigerian speaks English. While the Google Nigerian dataset comprises male and female speakers, SautiDB comprises the different tribal English accents across the country. This resulted in a total of 4,278 audio files.

• **Data preprocessing**: To develop an accurate ASR system, it is important to collect and preprocess the data in an appropriate form. The audio files were downsampled. This simply means lowering an audio signal's sampling rate. This is usually done to preserve memory. All audio files are downsampled to 16 kHz using the librosa library. The transcripts were converted to lower sentence cases, and all the punctuation was removed except for the apostrophe, which gives meaning to words.

• **ASR Model Development**: The experimentation in this work was performed leveraging NVIDIA NeMo Quartznet15x5 and Wav2vec2 pretrained ASR model

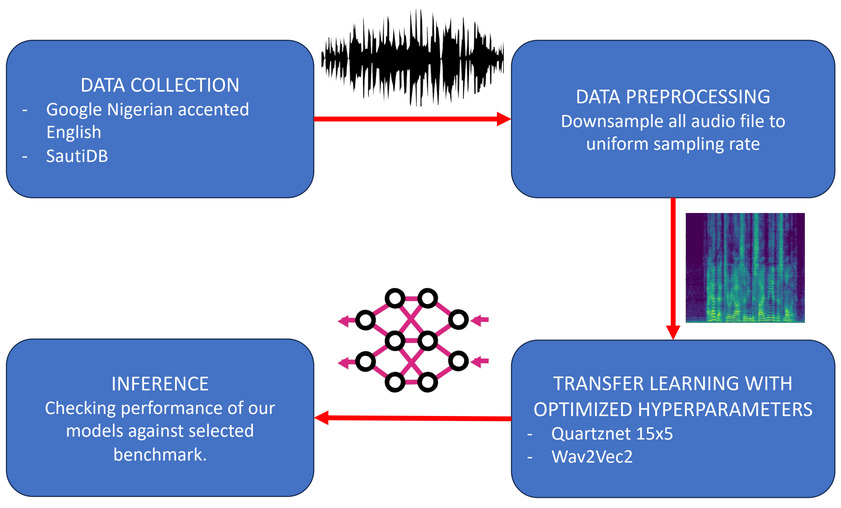

Figure 1. Our workflow

Results

The metric used to evaluate an ASR model is Word Error Rate (WER). WER is a metric commonly used to evaluate the performance of ASR systems. Upon comparing both models to the widely-used and free Google SpeechRecognition API using the Nigerian-English (en-NG) accent on the test data, we found that the results exhibited subpar performance with a test WER of 44.2%. The results of the two fine-tuned ASR models utilized in this paper, as shown in Table 1, indicate that Wav2Vec2, being a very big model, overfitted during the training process, resulting in a decline in performance on the validation and test datasets. In our experiment, NeMo QuartzNet15x5Base-En was found to be a better baseline for ASR model in the low data resource regime.

Table 1. Performance Summary Table of the ASR pretrained model on the African accented speech data.

Model	Train WER	Val WER	Test WER	Single Inference time	Training Duration
QuartzNet15x5	28%	17.6%	8.2%	0.156 secs	3h27m41s
Wav2vec2	19.7%	17.6%	14.9%	1.1 secs	7h21m48s

Conclusion

In this paper, we develop a Nigerian-accented English ASR system using a limited amount of labeled data from Nigerian English speech. We provided insights into the training and inference processes, highlighting the results and observations made. ASR for African-accented English is necessary to ensure inclusivity, effective communication, recognition, representation, and the advancement of natural language processing. By developing robust and accurate ASR systems that encompass diverse English accents, we create a more equitable and accessible technological landscape that respects and embraces linguistic diversity. This work advances NLP research and technology in recognizing poorly represented English accents and is intended to serve as a reference for future ASR research in the context of English accents.

References

[1] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689

[2] Dossou, B. F., Tonja, A. L., Emezue, C. C., Olatunji, T., Etori, N. A., Osei, S., ... & Singh, S. (2023). Adapting Pretrained ASR Models to Low-resource Clinical Speech using Epistemic Uncertainty-based Data Selection. arXiv preprint arXiv:2306.0210'.mn vjk,[]pm, ?/b5.

[3] Yemmene, P., & Besacier, L. (2019). Motivations, challenges, and perspectives for the development of an Automatic Speech Recognition System for the under-resourced Ngiemboon Language. In Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers (pp. 59-67).

[4] Ibejih, S., Oyewusi, W. F., Adekanmbi, O., & Osakuade, O. EDUSTT: In-Domain Speech Recognition for Nigerian Accented Educational Contents in English. In 3rd Workshop on African Natural Language Processing.

[5] Dossou, B. F., Tonja, A. L., Emezue, C. C., Olatunji, T., Etori, N. A., Osei, S., ... & Singh, S. (2023). Adapting Pretrained ASR Models to Low-resource Clinical Speech using Epistemic Uncertainty-based Data Selection. arXiv preprint arXiv:2306.02105.

[6] https://openslr.org/70/

[7] Afonja, T., Mudele, O., Orife, I., Dukor, K., Francis, L., Goodness, D., ... & Mbataku, C. (2021). Learning Nigerian accent embeddings from speech: preliminary results based on SautiDB-Naija corpus. arXiv preprint arXiv:2112.06199.

[8] Tamburini, F. (2021). Playing with NeMo for Building an Automatic Speech Recogniser for Italian. In CLiC-it.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural 283 Information Processing, 33:12449–12460, 2020

Table 2. Qualitative Comparison of the ASR pretrained model predictions

Actual Text	Transcriptions		
	Google SpeechRecognition Api	Quartznet15x5	Wav2vec2
the fula people or fulani are one of the largest ethnic groups in the sahel and west africa	the full of people are funny are one of the largest ethnic groups in the Sahel and West Africa	the fula people or fulani are one of the largest ethnic groups in the sahel and west africa	he fula people or fulani are one of the largest ethnic groups in the sahel and west africa
ade obayemi opined that the okun people are aboriginals in the niger benue confluence	Adele by me open people are aboriginals in the Niger benue confluence	ade obayemi opined that the okun people are aboriginals in the niger benue confluence	ade obayemi opined tat the okon people are oboriginals in the niger benue confluence
freyja says that loki is lying that he is just looking to blather about misdeeds	free just say that Luke is lying they just look into that about misdeeds	frado says that loki is lieing that is just looking to blaggter about mis steeds	fhredio says that lokiy s line dhey is just looking to blatter about mis deeds
gorgeously voluminous robes intricately embroidered are a symbol of prestige and rank for men in nupe and hausa communities	gorgeously voluminous robes intricately embroidered a symbol of prestige and rank for many nuclear and hausa communities	gorgeously voluminous robes intricately embroidered are a symbol of prestige and rank for men in uwe and hausa communities	gorgeously voluminous robes intricately embroidered are a symbol of prestige and rank for men in uwe and Hausa communities
kperogi was among the presidential speechwriters during obasanjo's administration	where would you was among the presidential speech writers during the passengers Administration	werogi was among the presidential speechwriters during abasajos administration	perogi was among the presidential speechwriters during obasanjos administration