Automatic Speech Recognition for Nigerian-Accented English

Oreoluwa Boluwatife Babatunde
Department of Mathematical Science,
(Computer Science),
Olabisi Onabanjo University
babatundeoreoluwa35@gmail.com

Emmanuel O. Akeweje
School of Computer Science and Statistics
Trinity College Dublin
eakeweje@gmail.com

Sharon Ibejih
sharonibejih@gmail.com

Victor Tolulope Olufemi
Department of Electronic and Electrical
Engineering, Obafemi Awolowo University
femiosinkolu@gmail.com

Sakinat Oluwabukonla Folorunso
Department of Mathematical Science,
(Computer Science),
Olabisi Onabanjo University
sakinat.folorunso@oouagoiwoye.edu.ng

Abstract

Automatic Speech Recognition (ASR) systems have become ubiquitous in our daily lives, powering voice assistants and transcription services. However, these systems often overlook the diverse range of accents, including Nigerian-accented English, as they are primarily developed and trained on native English accents. This research aims to address this gap by developing a Nigerian-accented English ASR system. By creating ASR models capable of accurately interpreting and transcribing Nigerian-accented English, we strive to ensure equitable access to ASR technologies and services for individuals with Nigerian accents. The study employed transfer learning techniques on NeMo's QuartzNet15x5 English model and Wav2vec2.0 XLS-R300M using Nigerian-accented data. NeMo QuartzNet15x5Base-En exhibited the fastest inference time of 0.156 seconds with a Word Error Rate (WER) of 8.2% on the test set and Wav2Vec2 XLS-R-300M achieved a WER of 14.9% on the test set with an inference time of 1.1 seconds. This work presents the NeMo QuartzNet15x5Base-En pretrain model as best for ASR modelling especially in a low-resource regime.

1 Introduction

African accents exhibit differences in pronunciation, vowel placements, mean fundamental frequency, and phone model distances in the acoustic space, among other factors, which contribute to the challenges faced by ASR systems in accurately recognizing African-accented English. These variations are often overlooked in the development and training of ASR models, resulting in biased and suboptimal performance [1]. 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

5th Deep Learning Indaba Conference (DLI 2023).

of African accents and dialects [2]. However, developing and deploying ASR systems in African settings, particularly in Sub-Saharan Africa, presents its own set of challenges. The diversity of African languages, with numerous dialects, accents, and variants, coupled with resource constraints in terms of finance, infrastructure, and technology, pose obstacles to ASR progress [3]. Overcoming these challenges requires innovative approaches, collaboration, and the effective utilization of limited data resources. 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.

2 Related works

In comparison to other languages, dialects, and accents, there has been limited research and improvement in the field of African-accented Automatic Speech Recognition (ASR) over the last three decades. [4] an Automatic Speech Recognition (ASR) system designed for educational purposes specifically targeting the Nigerian accent called EDUSTT. The audio data used in this study was sourced from a Nigerian learning platform that was created during a lockdown period to facilitate children's remote learning through radio broadcasts. The total voice data collected for this research consisted of 20 samples, amounting to a cumulative duration of 8 hours. The researchers conducted the experiment by fine-tuning NeMo's QuartzNet 15x5 English model using their educational data with Nigerian accents. The research by [5] focuses on Adapting Pretrained ASR Models to Low-resource Clinical Speech using Epistemic Uncertainty-based Data Selection. Using African-accented English datasets, including a custom clinical dataset (AfriSpeech200), they introduced an uncertainty-based algorithm for data selection and demonstrated that the algorithm is independent of the specific model utilized, with the most effective selection mode being used exclusively for these models because it has been proven to produce the greatest results.

3 Methodology

3.1 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.

3.2 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. The data was shuffled and split into the following:

Table 1: Dataset splitting and Duration

Data Set	Proportion %	Number of Samples	Duration(hours)
Train	70	3385	5.32
Validation	29	855	1.36
Test	1	38	0.076

3.3 ASR Model Development

The experimentation in this work was performed leveraging NVIDIA NeMo Quartznet15x5 and Wav2vec2 pretrained ASR model.

NeMo QuartzNet15x5Base-En: NeMo is a versatile Python toolkit for AI applications, providing reusability and composition capabilities. NVIDIA offers pre-trained ASR models, including the "STT en Quartznet15x5" model for English. NeMo is based on neural modules, which are the building blocks of neural networks that accept typed inputs and output typed outputs. Data layers, encoders, decoders, language models, loss functions, and techniques of mixing activations are common examples of such modules [8] NeMo's neural type system makes it simple to combine and reuse these building elements while also offering semantic correctness checking. The toolkit includes a set of pre-built modules for automatic speech recognition and natural language processing that can be expanded as needed. NeMo QuartzNet15x5Base-En was trained using over 3,300 hours of speech data from Mozilla's English Common Voice 6.1 and Multilingual LibriSpeech datasets. The training incorporated speed perturbation and Cutout data augmentation methods, and the Apex/Amp O1 optimization level was utilized. The model achieved a word error rate (WER) of 4.19% on the LibriSpeech test-clean dataset [4].

Wav2vec2.0 XLS-R300M: Wav2vec2 is an advancement in Facebook AI's research on pretraining large speech models using self-supervised learning, which can then be fine-tuned on smaller labeled data. This framework learns representations from raw audio data and utilizes a multi-layer convolutional neural network for encoding speech [9]. It incorporates masked language modeling to mask latent speech representations.Wav2vec2.0 XLS-R300M has 300 million parameters . Results shows that learning discrete speech units and contextualized representations simultaneously outperforms previous methods that use fixed units. Using only 10 minutes of labeled data, this approach achieves a word error rate (WER) of 4.8/8.2 on the clean/other test sets of Librispeech, demonstrating the potential of ultra-low resource speech recognition [9].

3.4 Model Training

As an experiment, the training was kept simple with some fixed hyperparameter values for equal performance evaluation. The batch size used for finetuning all pretrained models was 8, and the number of epochs was 50. The training ran on Nvidia Tesla P100 gpu.

4 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.

Table 2: 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

Both NeMo QuartzNet15x5Base-En and Wav2vec2.0 XLS-R300M have demonstrated their robustness and effectiveness in handling small data, as shown in Table 3. 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 2, 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. From our experiment, NeMo QuartzNet15x5Base-En was found to be a better baseline for ASR model in the low data resource regime.

Table 3: 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

5 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 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 researches 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:2306.02105.

[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