EDUSTT: In-Domain Speech Recognition for Nigerian Accented Educational Contents in English

Sharon Ibejih, Wuraola Fisayo Oyewusi, Olubayo Adekanmbi & Opeyemi Osakuade
Data Science Nigeria (Data Scientists Network)
Lagos, Nigeria.
{sharon, wuraola, olubayo, osakuade}@datasciencenigeria.ai

Abstract

English Automatic Speech Recognition systems are trained on regular speech, therefore they may struggle to perform well on accented and domain-specific speech. For broader applications of ASR systems, such as in education, where there is synchronous learning, it is important to have a reliable system - a specialized system that recognises terms used in school subjects and spoken by accented teachers. English is our official language in Nigeria, and it is the major language used to teach in schools. However, our teachers hail from different parts of the country, where their mother-tongue affects the way they pronounce certain words. The aim of this paper is to propose an ASR system for education in Nigerian accent. Our experiment leveraged on fine tuning NeMo's QuartzNet15x5 English model on our accented educational data. This process yielded a WER of 27%.

1 Background

Automatic speech recognition (ASR) is a language technology that recognizes, interprets, and converts spoken words to text Dossou & Emezue (2021). Speech recognition has been a progressive technology over the last thirty years and currently, we have systems that are quite efficient in transcribing spoken words. There are limitations in the use of ASR technologies but over time, this technology has proven to be beneficial in education by striving to overcome the limits of physical and online synchronous learning environments. Students can use generated texts to better understand lecture content, validate missing or misheard sections of a speech, take notes or complete assignments, and prepare for exams Shadiev et al. (2014). It has been proposed that speech recognition generated texts can considerably assist students in better understanding lectures, taking simultaneous notes during lectures, and completing assignments Shadiev et al. (2014)Hwang et al. (2010)Kuo et al. (2012). Furthermore, it is believed that ASR-generated text can be used as an additional text confirmation of what is being said, and that it can assist comprehension in situations when listeners are students with learning or physical challenges, foreign students, or other at-risk populations Kuo et al. (2012)Wald & Bain (2008).

However, English speakers have a wide range of accents, which are used in both English (literature and art disciplines) and non-English terms in audio instructional content (like Biological or Mathematical subjects or native names in a History subject). As a result, existing ASR systems in educational applications do not have a high level of accuracy. When students began to learn electronically in the year 2020, during the COVID-19 pandemic, it became clear that English speech recognition algorithms failed to correctly transcribe the pronunciations of educators with Nigerian accents, as well as academic concepts that were not proper English words. Streamed classes on YouTube, for example, were the subject of our case study. To avoid erroneous transcriptions, pupils have to rely only on audio content. When you consider the difficulties of virtual learning, such as a poor network connection, depending solely on audio information has negative influence on students' learning. The goal of this study is to address this problem and develop an end to end speech recognizer for educational content with Nigerian accents.

1.1 Overview of Nigerian Accent in Educational Content

The English language is spoken practically everywhere in the world, including in Nigeria, where it is used for governmental reasons. It is the teaching language in our schools, and it is highly regarded among its speakers in Nigeria. As a multicultural country with over 200 languages, the accents with which locals speak English and pronounce educational and technical phrases, are sometimes influenced by their cultural languages' distinct tones. It invariably affects understanding for people who do not speak the same language or have a similar accent.

There are three primary ethnic groupings in the country - Igbo (South-Eastern zone), Yoruba (South-Western zone), and Hausa (Northern zone). While native Igbos are recognized for interchanging "l" and "r", like saying "lemove" instead of "remove", the Yorubas are known for pronouncing "h" before any vowel sound and suppressing h in words where it appears. Saying "ouse" instead of "house" and "heat" instead of "eat" are two examples. Hausa, on the other hand, is a tonal language in which different words and expressions have varying tones. All of these distinguishable phonologies are influenced by the construct of the respective native languages.

At some educated levels, the vowel systems of these three major Nigerian varieties are said to coalesce. They come in two flavors: one geared toward standard dialect and the other geared toward Igbo-Yoruba English. Most Hausa English speakers, particularly in Katsina, Kaduna, Nassarawa, Kano, Jigawa, Sokoto, and Kebbi states, have a system that is similar to that of basic, sophisticated, or educated speakers Olaniyi & Josiah (2013). Some educated people who have mastered English vocalists, like broadcasters, may have good experience with ASR systems because of the clarity in their English pronunciations compared to an average educated Nigerian whose English accent is still affected by their mother tongue.

1.2 Related Work

The majority of research focuses on speech recognition for Nigerian local languages, which is not our project scope. There is no documented work on in-domain speech recognition for education in Nigerian English accent as of this writing. Therefore, we find this work interesting as it explores the potential opportunity of ASR on the education threshold especially with accented teachers.

We explored the use of a base model that had been trained on around 7,000 hours of English speech to fine-tune on our data. The approach used is the NVIDIA NeMo ASR QuartzNet15x5 English Model.

2 Method

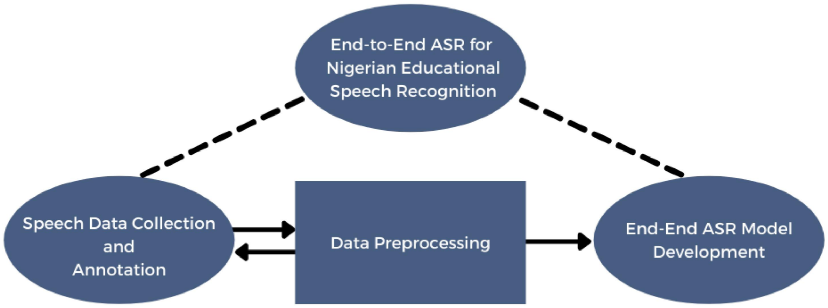
End-to-End ASR for Nigerian Educational Speech Recognition
Speech Data Collection and Annotation
Data Preprocessing
End-End ASR Model Development

Figure 1: Research Methodology

2.1 Speech Data Collection and Annotation

2.1.1 Data Collection

Our audio data came from a Nigerian learning platform that our organization established during the lockdown to assist children learn at home through radio broadcasts. The information is in the form

Table 1: Subjects, number of files and duration of educational content used in training in EDUSTT

Subjects	Number of files	Duration (Hours)
Mathematics	6	2.7
English	5	2
Basic Science	5	2
Biology	3	1.2
Economics	1	0.4

of audio learning materials and exam preparation tools created by consensual teachers based on the recognized national curriculum in the fields of Mathematics, English, Economics, Biology, and Basic Science. The instructors are men and women from all across the country who speak English with a variety of accents. The total amount of voice data was 20, for a total time of 8 hours, which was then divided into train, validation, and test sets.

2.1.2 Data Preparation and Transcription

Each of the 20 audio files is 24 minutes long. Using the Audacity® software [1], they were all chunked into 5 seconds and exported in WAV format. This resulted to 5685 speech samples. Each speech sample averagely made up ten words and were manually annotated (transcribed). The annotators spent about 4 hours transcribing 300 seconds of audio. We made use of a text editor to write down pronounced words asynchronously, including stammers, and ehms, to prevent out of vocabulary issues.

Rules that guided the annotation:

1. Pronunciations influenced by accents were transcribed with their correct spellings. For example, transcribing "hion" as "ion".
2. All numbers were transcribed in words. For example, JSS two, instead of “JSS 2”.
3. Word spellings in the audios were transcribed as individual letters. For example, g a r b a g e
4. The truncated pronunciation at the end of each speech sample was also transcribed based on the fraction that was heard. For example, “sch”, a truncated part of “school”.
5. Words were spelt in full, e.g “Doctor” instead of “Dr”
6. Punctuation marks were not used except for apostrophes because writing “Im” instead of “I’m” could be meaningless.
7. Using "[unknown]" to represent inaudible or unclear pronunciations.
8. After transcribing, the name of the audio file is written in a bracket just next to the transcription for easy data preprocessing.

2.2 Data Preprocessing

The audio samples were preprocessed according to our model development technique’s specifications. For computation, all signals were downsampled to 16000Hz from a sample rate of 44100Hz. It was also crucial to double-check if the audios had a mono channel. The train and validation sets were then converted to manifest files. The manifest files are simply in JSON format containing the audio file path of each sample, their durations and transcriptions.

2.3 End-to-End ASR Development

2.3.1 NeMo Toolkit

NeMo (Neural Modules) is an open-source Python framework-agnostic toolkit for creating AI applications through re-usability, abstraction, and composition. NeMo is built around neural modules, conceptual blocks of neural networks that take typed inputs and produce typed outputs. Such modules typically represent data layers, encoders, decoders, language models, loss functions, or methods of combining activations. NeMo makes it easy to combine and re-use these building blocks while

[1] Audacity® software is copyright © 1999-2021 Audacity Team. The name Audacity® is a registered trademark.

providing a level of semantic correctness checking via its neural type system. The toolkit comes with extendable collections of pre-built modules for automatic speech recognition and natural language processing Kuchaiev et al. (2019).

2.3.2 MODEL DEVELOPMENT

This project adopted NeMo's pretrained ASR QuartzNet15x5 English model for fine-tuning on our Nigerian accented educational data. QuartzNet [2] family of models are denoted as "QuartzNet [BxR]", The model's internal structure is displayed in 2, where B is the number of blocks and R is the number of convolutional sub-blocks within a block. The convolutional model trained with Connectionist Temporal Classification (CTC) loss, is designed for tasks where we need alignment between sequences, but where that alignment is difficult - e.g. aligning each character to its location in an audio file. It calculates a loss between a continuous (unsegmented) time series and a target sequence.

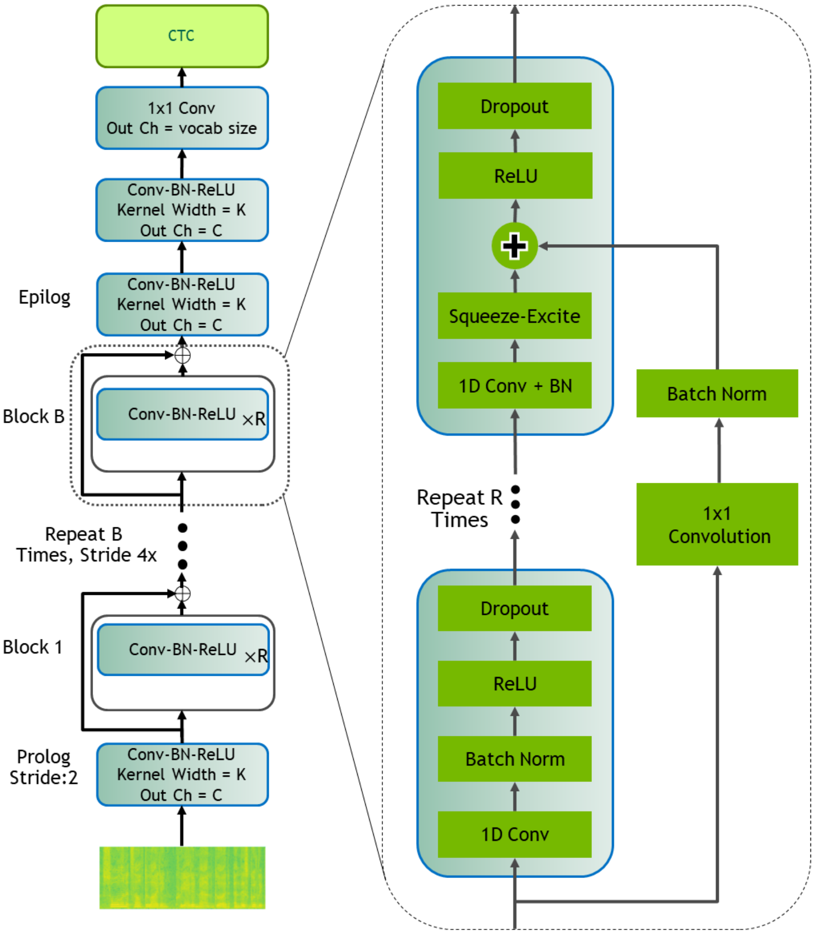

Figure 2: Quartznet BxR model with 1D Time-Separable Convolutions

The description says QuartzNet15x5 model trained on six datasets: LibriSpeech, Mozilla Common Voice (validated clips from en_1488h_2019-12-10), WSJ, Fisher, Switchboard, and NSC Singapore English. It was trained with Apex/Amp optimization level O1 for 600 epochs. The model achieves a WER of 3.79% on LibriSpeech dev-clean, and a WER of 10.05% on dev-other Shmyrev (2021).

[2] https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/asr/models.htmlquartznet

Table 2: WER Comparison across Models

Data Splits	Duration (Hours)	Number of samples
Train	7.79	5615
Validation	0.08	60
Test	0.035	25

The implementation technique involved changing the model's vocab to include more characters like apostrophe (""""), space(" ") and square brackets ("[", "]"). This change only affects the model's decoder. The encoder where the weights are stored, remains intact. The model was then trained on our data for 100 epochs with a learning rate of 0.001, using GPU from Google Colaboratory. Our work progressed from an initial training on an audio duration of 3.82 hours to 7.79 hours, being what this paper is presenting.

3 Results

Word Error Rate (WER) is a metric used to compute the similarity between a reference text and a prediction. It is the percentage of errors for every 100 words. It is computed as the ratio of the numerator and denominator for each pair. The lower the WER the better.

$WER = (S + D + I)/N$

where:
S is the number of substitutions
D is the number of deletions
I is the number of insertions
C is the number of correct words
N is the number of words in the reference, $N = S + D + C$.

The initial training with 3.82 hours speech corpus, generated a WER of 33%, while an increase in the training data size to 7.79, reduced the WER to **27%**. We selected it for a qualitative comparison with QuartzNet15x5 model and the two models fine tuned with 3.82 and 7.79 hours. This is shown in Table 3. The comparison is carried out on 6 test speech samples containing different school subjects. The words in bold are used to identify wrong transcriptions.

3.1 Discussion

In this work, we explored the working of a bespoke ASR model that is useful for different school subjects like Mathematics, English, Biology and Economics taught by accented teachers. The results in Table 3, prove that while there are good commercial speech recognition API with low WER, they are not very robust to accented English speakers. We tested on subjects that the model was trained on and subjects it has never seen such as Physics and Geography, to check for robustness in terms of accent. While, our model (7.79 hours) struggled with transcribing certain words, it makes its prediction based on phonemes, unlike the Google API that misinterpreted and completely omitted some pronunciations it didn't understand because it prioritizes predicting correct vocabularies over phonemes. For example, in the Biology sample, we have "stomac" for "stomach" and Google has "Visa" for "these".

A comparison between the models fine tuned on 3.82 hours of data and 7.79 is evident that more hours of data will boost the performance and robustness of our model in terms of educational content and accent.

Table 3: Transcription comparison of ASR models on test audio samples

Reference	**Google SR API**	**QuartzNet15x5 model**	**3.82 speech hours**	**7.79 speech hours**
PHYSICS: one or more of alpha particles beta particles and gamma-rays alongside energy are	one or more of a **whole bunch of Jews** and gamma-rays alongside energy **–**	one or more of **apanticles beaser** particles and **gammorays** and **ongside** energy **–**	**on** or more of **or** particles **esor** particles and **gamolraces andounside** energy are	**onone** or **urm** of **our** particles **leaser** particles and gamma rays and **oncide** energy are e
BIOLOGY: pale body or skin color swollen stomach and also thin limbs and legs these are characteristic feature	skin color swollen stomach and also thin **liens** and legs **Visa** characteristic feature	body or skin **colar** swollen stomach and also thin **leams** and legs these are characteristic features	**lbody** or skin **colar swolen** stomachs and also thin **lins** and **legks** these are characteristic feature	pale body or **skan** color **sfallow stomac** and also thin **mens** and legs these are characteristic **fetre**
ECONOMICS: A frequency B mean C median D mode I'll re	frequency mean median mode	a frequency b **mil s** median d **mod arit**	a frequency be mean see medium the mode **iw'l**	a frequenc b mean c **media** d **mood** ilre
GEOGRAPHY: chemical weathering solution oxidation carbonation hydration and hydrolysis	chemical weathering solution oxidation carbonation hydration and hydrolysis	**egin r** solution oxidation **cambonation** hydration and **idylysis**	**elial werding** are solution **axidation corbination** hydration and **ajolisis**	**ilica wardin** are solution oxidation **cobonation** hydration and **igalysis**
ENGLISH: r today's lesson so please take away all forms of distraction and lets have your one hundred percent focus that's good guys so let's begin today's class now like i said that today our topic is what identification of mood through the use of words and expression now at the end of this particular lesson you are expected to be able to identify the author's mood especially	take away all forms of distractions and let's have you one hundred percent Focus especially ***[...the model missed recognising a huge part of the audio]***	to these **mesons** so please take away all forms of **destructions** and let's have your one hundred per cent **forcas** that's good guy **ciles** begin to this class now like i said that they **artapecki's** word's identification **will move** through the use of words and expression now at the end of this **partcularlerson youare** expected to be able to identify the **otther's** mood especially	r todays a **essen** so please take away all forms of distraction and lets have your one hundred **persen** focus that **cold giso** let's begin **to dhays** class now like i said that today our topic is what identification of **move** through the use of words and expression now at the end of this particular lesson you are expected to be able to identify the authors mood especially	**ret** today's **nesson** so please take away all **foms** of distractions and lets have your one hundred percent focus that good **gays** so lets begin today's class now like i said that today our topic is what identification of **move** through the use of words and expression now at the end of this particular lesson you re expected to be able to identify the authors mood especially
MATHEMATICS: I want to start this way now what is a Factor let's start from there what is a factor you must have heard factor factor factor okay so what is a factor simply put simply put it this way a factor is a number that divides a given number without leaving any remainder	I want to stop this way now what is it Factor let's start from there what is a factor you must have it **faxed** a factor factor okay so what's the factor simply **boot** simply put it this way if factor is a number that divides a given number without leaving any remainder	i want to start this way now what is a factor let's stat from there what is a **fattor** you must have had **fanzo fa tofato oky** so what's a factor simply put simply **pitted** this way if factor is a number that divides a given number without leaving any remainder i	i want to start this way now what is a factor let's start from there what is a factor you must have heard **facof far top** factor okay so whatis a factor simply put simply **putted thas** way a factor is a number that divides a given number without leaving any reminder	i want to start this way now what is a factor lets start from there what is a factor you must have **eard fat far thouh** factor okay so what a factor simply put simply put it **thas** way a factor is a number that divides a given number without leaving any reminder i

4 Conclusion

This work presented EDUSTT, a bespoke ASR model for educational content in Nigerian English accent. It is part of an ongoing research and the current traction has shown usefulness for continued work. This model is openly available for use.

References

Bonaventure FP Dossou and Chris C Emezue. Okwugb\'e: End-to-end speech recognition for fon and igbo. *arXiv preprint arXiv:2103.07762*, 2021.

Wu-Yuin Hwang, Rustam Shadiev, Chiu-Tien Kuo, and Nian-Shing Chen. A study of speech to text recognition and its effect to synchronous learning. 06 2010. ISBN 978-1-880094-81-5.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan Cohen. Nemo: a toolkit for building ai applications using neural modules, 09 2019.

Chiu-Tien Kuo, Rustam Shadiev, Wu-Yuin Hwang, and Nian-Shing Chen. Effects of applying str for group learning activities on learning performance in a synchronous cyber classroom. *Computers Education*, 58:600–608, 01 2012. doi: 10.1109/ICALT.2011.74.

Oladimeji Olaniyi and Ubong Josiah. Nigerian accents of english in the context of world englishes. *World Journal of English Language*, 3, 01 2013. doi: 10.5430/wjel.v3n1p38.

Rustam Shadiev, Wu-Yuin Hwang, Nian-Shing Chen, and Yueh-Min Huang. Review of speech-to-text recognition technology for enhancing learning. *Educational Technology Society*, 17:65–84, 11 2014.

Nickolay Shmyrev. Nvidia nemo test results, 02 2021.

Mike Wald and Keith Bain. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, 6(4):435–447, 2008.