Nollywood: Let's Go to the Movies!

John E. Ortega *, William Chen, Ibrahim Said Ahmad

Posted Date: 15 February 2024

doi: 10.20944/preprints202402.0845.v1

Keywords: Natural Language Processing; Automatic Speech Recognition; Machine Translation; Nigeria; English; United States; Movies; machine learning

 doi:10.20944/preprints202402.0845.v1

Article

Nollywood: Let's Go to the Movies!

John E. Ortega [1,†], William Chen [2,‡] and Ibrahim Said Ahmad [3,†]

[1] Northeastern University; j.ortega@northeastern.edu
[2] Carnegie Mellon University; williamchen@cmu.edu
[3] Northeastern University; i.ahmad@northeastern.edu
* Correspondence: j.ortega@northeastern.edu

Abstract: *Nollywood*, based on the idea of Bollywood from India, is a series of outstanding movies that originate from Nigeria. Unfortunately, while the movies are in English, they are hard to understand for many native speakers due to the dialect of English that is spoken. In this article, we accomplish two goals: (1) create a phonetic sub-title model that is able to translate Nigerian English speech to American English and (2) use the most advanced toxicity detectors to discover how toxic the speech is. Our aim is to highlight the text in these videos which is often times ignored for lack of dialectal understanding due the fact that many people in Nigeria speak a native language like Hausa at home.

Keywords: natural language processing; automatic speech recognition; machine translation; machine learning; Nigeria; English; United States; movies

1. Introduction

In the past several decades, there has been a significant amount of research on digital systems pertaining to language. Some state-of-the-art digital language systems, like those based on automatic speech recognition (ASR), are now considered to be on par with humans for high-resource languages like English and Spanish for conversational speech recognition [1]. However, there are still challenges that significantly impact the performance of ASR, such as the recognition of English with accents [2]. The difficulty in recognizing accented English can be attributed to the diversity in pronunciation, intonation speed, and pronunciation of specific syllables.

In some countries, low-resource languages can affect how the high-resource language is spoken which poses many challenges from the ASR standpoint of view. For example, one recent study [3] shows that more often than not, ASR systems can be non-inclusive. The struggles that one may encounter as an end user are not the focus of this article but provide a backdrop for a problem that we implicitly address: how does the culture and other attributes of an English speaker affect the digital processing of speech and other resources. In this article, we focus on two main sub-fields of digital processing: *speech recognition* (SR) and *toxicity* (TX).

For our study to be valid and useful for those who speak low-resource languages like Hausa (a low-resource language spoken in Nigeria, Africa), we dive into a Nigerian digital movie genre called: **Nollywood**. Nollywood is the Nigerian video film industry that emerged within the context of several pre-existing theatre traditions among various ethnic groups of Nigeria [4]. Nollywood is currently the third largest film industry in the world, and it has generated over 500 million dollars since inception [5]. According to [6], Nollywood is gaining popularity due to its transformations, which bear resemblances to the processes of gentrification and professionalization. This is formalizing the industry as well as attracting professionals and instigating existing filmmakers to improve on their art. Nollywood has impact not just in Nigeria, but across the African continent and beyond.

In this paper, we focus on ASR for two countries where English is the official (high-resource) language: Nigeria and the United States of American (USA). These two countries represent a large part of the English-speaking world. Nigeria has a population of more than 200 million[1] while the

[1] https://www.census.gov/popclock/world/ni

 doi:10.20944/preprints202402.0845.v1

USA has a population of more than 300 million[2]. Both countries generally use English for business and everyday conversation since their official languages are English [7]. However, in areas such as New York City in the USA and in most of Nigeria, native-language accents can have an influence on how English is spoken. It is a known fact that ASR systems are known to be faulty when tested with non-native speakers [8].

The authors of this work all speak English as it is the high-resource language of their countries. As one of the co-authors speaks Hausa and has first-hand experience, we consider our approach somewhat more inclusive. Moreover, this article brings attention to specific dialectical approaches that could be used in other languages such as Spanish or Chinese (languages which the other two authors speak fluently). We feel that this gives us insights into the problem that other investigators may not have. There is a difference between the accent influence in Nigeria versus that of the USA. In our opinion, English takes the forefront of culture in the USA and has a dominant force, especially in the movie industry. That does not seem to be the same in Nigerian movies such as the Nollywood movies. In this article, we attempt to build a state-of-the-art ASR system for Nigerian English to see if the system performs well when applied to the Nigerian dialect. Additionally, we compare the two languages to determine the amount of toxicity (words that are considered taboo or should not be spoken in formal language).

In order to better describe our experimentation, we divide this article into several sections. First, in Section 2 we introduce work that has motivated our experiments and is related to ours. We then cover our approach in more detail for ASR and TX in Section 3. Thirdly, we compare our results in Section 4. Finally, we provide further analysis and conclude in Section 5 and provide some ideas of future work in Seciton 6.

2. Related Work

The article by Amuda *et al.* [9] investigates the engineering analysis and recognition of Nigerian English (NE) in comparison to American English (AE) using the UILSpeech corpus. The study utilizes speech audio and video data to analyze speech parameters and their impact on automatic speech recognition (ASR) systems. Data collection includes isolated word recordings and continuous read speech data from Nigerian English speakers, highlighting the linguistic diversity of the region. The research employs techniques such as lexicon extension and acoustic model adaptation to address phonetic and acoustic mismatches between NE and AE, resulting in a 37% absolute word error rate reduction. The main findings emphasize the importance of tailored lexicons and acoustic models for low resource languages, showcasing the potential for improved ASR performance in dialectal variations. However, the study acknowledges limitations in the generalizability of findings to other low resource dialects and the need for further research on speaker-dependent phonetic patterns.

Babatunde [10] presents a novel approach to developing an Automatic Speech Recognition (ASR) system tailored for Nigerian-accented English by leveraging transfer learning techniques on pretrained models, including NeMo QuartzNet15x5 and Wav2vec2.0 XLS-R300M. The research addresses the challenges of recognizing and transcribing Nigerian accents, aiming to ensure equitable access to ASR technologies for individuals with Nigerian accents. The NeMo QuartzNet15x5Base-En model demonstrated promising results with a Word Error Rate (WER) of 8.2% on the test set, showcasing its effectiveness in handling Nigerian-accented speech data. However, limitations such as the small dataset size and overfitting observed in the Wav2vec2.0 XLS-R300M model were acknowledged. This work contributes to the advancement of ASR systems for African-accented English, emphasizing the importance of inclusivity and accurate transcription in diverse linguistic communities.

Oluwatomiyin *et al.* [11] explores the development of a hybrid translation model for converting pidgin English to the English language, utilizing the JW300 corpus for training and evaluation. The

[2] https://en.wikipedia.org/wiki/Demographics_of_the_United_States

 doi:10.20944/preprints202402.0845.v1

study employs Phrase-based Statistical Machine Translation (PBSMT) and Transformer-based Neural Machine Translation models to enhance translation accuracy. Results indicate that the hybrid model surpasses the baseline NMT model, demonstrating improved performance in translation tasks with the highest BLEU score of 29.43 using the pidgin-pbsmt model. The findings highlight the potential of combining PBSMT and NMT techniques to enhance translation quality for low-resource languages. However, limitations include challenges related to vocabulary size and computational resources, suggesting the need for further research to address scalability issues and optimize the model for broader applications.

An article Oladipupo and Akinfenwa [12] examines the phonemic realisation of educated Nollywood artistes in Nigeria and their accent as a normative standard of English pronunciation. It analyzes the pronunciation of various phonemes, comparing them to Received Pronunciation (RP) forms. The study focuses on the competence of educated Nollywood artistes in pronouncing these phonemes, highlighting improvements in the realisation of certain vowels compared to typical Nigerian English accents. The research suggests that these artistes could serve as normative pronunciation models for Nigerian English learners. Additionally, the article discusses the debate surrounding the codification of Nigerian English as a standard for communication and learning in Nigeria, considering the influence of native English models and technology-driven speech practice sources.

3. Methodology

Our experiments represent two of the most modern tasks currently being investigated in the natural language processing (NLP) field. Recent conferences such as Interspeech[3] and ACL[4] have including ASR and TX as main focuses in workshops and other publications. While low-resource languages are heavily investigated for tasks such as machine translation (MT), the effect of low-resource language speaker's accents on English is not heavily researched. In this section, we present several state-of-the-art systems for ASR and TX with a focus on dialectal difference between English spoken in movies from Nollywood and Hollywood.

In order to clearly illustrate the need to better identify the difference between the two dialects, we first motivate our experiments visually by comparing identical sentences spoken by two male counterparts: a Nigerian speaker and a USA speaker. We use these examples as ample evidence to show that there is quite a bit of difference between the two for sentences that are not complex. In Figure 1, we compare and contrast spectrogram samples for the follow four sentences.

Sentence 1:
Hey, have you heard the latest gist about the party next weekend? It's gonna be lit!
Sentence 2:
Let's schedule the meeting for October 10th at 2:30 PM.
Sentence 3:
I'll meet you at the gas station; we can take the freeway to the shopping mall.
Sentence 4:
The project deadline is tomorrow, and I need to submit my resume to the recruiter.

[3] https://interspeech2023.org/
[4] https://2023.aclweb.org/

 doi:10.20944/preprints202402.0845.v1

In order to create the spectrograms, we had to find a way to create an audio file (.wav format) that would take as input one of the four sentences from Figure 1. An online tool called SpeechGen[5] allowed us to perform the task and the output was verified by the authors, Nigerian and USA native speakers.

It is clear that there is a noted difference visually from the generated speech files for Sentences 1 through 4 in Figure 1. For example, for Sentence 3 the US English (e) and Nigerian English (f) between seconds 2.0 and 2.5 are quite different. The Nigerian speaker seems to have higher frequency and contain more volume. While the audio files where created using a digital ASR system; actual human voice could be more expressive. In this effort, we wanted the system to be equal in order to measure the main digital difference between the two languages as no two humans can be considered to have the same dialect or voice [13]. With the notion of difference between the two dialects we present the following steps taken to repeat our experiments.

3.1. Corpora

For our experiments to be realistic and capture differences in everyday movies, we assess two major films: (1) the movie Deep Cut[6] from the Nigerian Nollywood theme and (2) Acrimony[7] from the US Hollywood them. We use these two examples as random picks for movies that could be considered from the nearly the same genres and containing similar topics. We gather the text transcriptions from both movies to assess first *toxicity*. Then, for a more direct evaluation of how well ASR fares on Nigerian corpus, we use the Nigerian ICE corpus [14]. We did not test on ASR for the USA dialect as it has already been reported on in several conferences.

3.2. Toxicity

Toxicity detection is an important challenge in NLP as the latest research [15,16] shows. Frameworks, like the MT one from Meta called Seamless4MT [17], use modern techniques such as large language models and other generative techniques to resolve several issues in translation. We consider the framework state-of-the-art for translation and stable enough to be considered a good determiner for finding toxicity in text. In our experiments, we mirror Meta's work by using ETOX[8], a pre-trained model found on Hugging Face[9] for toxicity detection. For comparison purposes, we also use another common framework called Evaluate[10] used to measure bias.

We measure the prevalence of toxicity in the two films (Deepcut and Acrimony) as an initial manner to help those working on dialect detection in Nigerian better understand if words from the two movies were found to be more toxic in one dialect of the other. While the comparison is not an identical comparison, we were unable to use SpeechGen (the ASR generation tool used for creating spectrograms from Figure 1) to create a corpus to measure identical movies.

3.3. Automatic Speech Recognition

ASR tools are plentiful for USA English; however, here our main goal is to test the validity of the latest technique for a dialect in English that is not in mainstream research: Nigerian. In order to do that, we use the largest corpus we could find for training an ASR model on English text with a Nigerian accent: **ICE**.

In order to test ICE, we experiment with two ASR models that are widely used as novel models at this point in time: *Whisper* [18] and *XLS-R* [18]. Both models have been found to perform well on recent speech tasks such as the International Workshop on Speech Translation 2023 [19]. To be more

---

5 https://speechgen.io
6 https://www.youtube.com/watch?v=Xl6ANUHjEtI
7 https://en.wikipedia.org/wiki/Acrimony_(film)
8 https://github.com/facebookresearch/stopes/tree/main/demo/toxicity-alti-hb/ETOX
9 https://huggingface.co/spaces/evaluate-measurement/toxicity
10 https://huggingface.co/blog/evaluating-llm-bias

 doi:10.20944/preprints202402.0845.v1

specific, we use an augmented version of XLS-R that was fine-tuned on multiple languages [20] with the hope that it may capture dialect differences. In order to fine-tune the model we use 22 hours of randomly-selected audio files (.wav) from the ICE corpus. As a form of validation/development we used 7.5 hours of files and tested on 9 hours. For Whisper, since it is generally used for what is known as *zero-shot* recognition and does not require fine tuning, we use the latest version (commit 1838) from OpenAI[11].

4. Results

Our results are dividing into the two tasks presented: Toxicity and ASR. We present our findings with the corresponding metrics. For toxicity, measurements are done using a percentage from 1 to 100% where 100% represents full toxicity. For ASR, we use the standard metric known as word-error rate (WER) which measures the number of words correctly predicted by the model.

In Table 1, we provide the results of the ETOX toxic evaluation tool on both movies along with the ICE corpus. Toxicity is on par for both languages as expected. While both movies are related to family topics, we do not have a parental rating for Deep Cut. Acrimony is rated 'R' by the USA administration; therefore, it can be expected to have somewhat more toxic language. Additionally, other factors like diversity in the USA[12] that mark differences between the two countries *could be* considered as important factors. However, it is out of the scope of this paper and saved for future work.

Measurements for the ICE corpus across all sets: *train*, *development/validation*, and *test* were below 1% and statistically insignificant. As part of our next iteration, we would like to consider more corpora of different genres and dialects.

Table 1. Toxicity results.

	Deepcut	Acrimony	ICE Spoken	ICE Written
ETOX	2.08%	3.35%	<1%	<1%
Evaluate	1.30%	2.16%	<1%	<1%

ASR for Nigerian was remarkably insufficient using the latest techniques. In some cases, the amount of text produced by what can be considered novel techniques introduced words that had no match, causing WERs higher than 100%. Results for the Whisper and XLS-R approaches are found in Table 2.

Our experiments show that for the Nigerian Deep Cut movie, WER in excess of 100% (124 and 231 for Whisper and XLS-R respectively). Our analysis shows that the multilingual nature of Whisper seems to produce words for the Nigerian English speech into another language spoken in Africa such as Arabic or even Devanagari, a language common in Northern India. At this point in time, we do not have an explanation of why Whisper produces this type of text, one thought is that the dialects from Arabic and Devanagari may be somewhat similar to the Nigerian dialect – at this point we are clearly assuming and leave verification of intuition for future work.

Whisper fails to recognize the Nigerian speech, with a WER of over 90%. We found that this is usually because Whisper is unable to properly identify the language being spoken, often incorrectly transcribing the speech into Arabic or Devanagari text. On the other hand, the XLS-R model, despite it being fine-tuned on Nigerian speech, does not perform well either. We believe that this could be due to the lack of Nigerian English in the XLS-R training data. For the ICE corpus, on the other hand, Whisper performs under 100% with nearly 94% error which is considered to be remarkably erroneous when compared with other English ASR systems like those created for USA English. XLS-R contrastingly

[11] https://github.com/openai/whisper
[12] https://www.census.gov/newsroom/blogs/random-samplings/2023/05/racial-ethnic-diversity-adults-children.html

 doi:10.20944/preprints202402.0845.v1

scores quite well compared to all of the other ASR systems with about 40% WER on the ICE corpus. We consider these findings important and feel that further hyper-parameter search using XLS-R is warranted.

Table 2. Automatic Speech Recognition results by using word-error rate (↓)

	Deepcut	**ICE**
Whisper Small	123.5	93.8
XLS-R	230.8	39.9

5. Conclusions

We conclude our experiments and findings with a clear explanation: *Nollywood movies are great to watch but hard to process*. The experiments performed show that, despite the great advancements in English, the high-resource language used more often for experiments in Artificial Intelligence, low-resource language influence on languages like Nigerian make it more complex to process and build tools for such nations. It is comforting to know that the Nollywood movie along with the formal ICE corpus seem to be less biased and contain less toxicity than their USA counterparts.

The goal of this paper was to show that Nollywood movies from Nigeria should be considered high-quality movies to watch. Although, if one would like to watch them in their dialect (allbeit English), it may be a while as research has not been advanced much in this area. Additionally, while several African languages like Tamasheq and others are becoming more prevalent in large tasks such as IWSLT 2023 [19], dialectal tasks should include other dialects such as the English dialect from Nigeria.

6. Future Work

We have noted through this article several investigative opportunities which we feel need to be addressed. For example, this work focuses on two mainstream movies and corpora. The next step would be to perform a large-scale search and inclusion of Nigerian corpora. Toxicity and bias can be measured on those corpora and should be compared to more corpora from the USA or other countries. Our systems generated words in Arabic and Devanagari. The next investigations should be to better understand why these systems produce those words for English spoken with a Nigerian dialect.

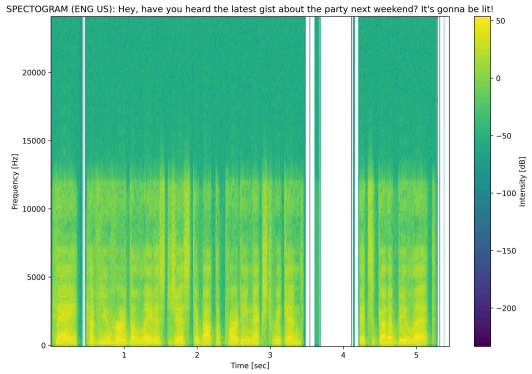
References

1. Min, Z.; Wang, J. Exploring the Integration of Large Language Models into Automatic Speech Recognition Systems: An Empirical Study. Neural Information Processing; Luo, B.; Cheng, L.; Wu, Z.G.; Li, H.; Li, C., Eds.; Springer Nature Singapore: Singapore, 2024; pp. 69–84.
2. Hinsvark, A.; Delworth, N.; Rio, M.; McNamara, Q.; Dong, J.; Westerman, R.; Huang, M.; Palakapilly, J.; Drexler, J.; Pirkin, I.A.; Bhandari, N.; Jette, M. Accented Speech Recognition: A Survey. *ArXiv* **2021**, *abs/2104.10747*.
3. Ngueajio, M.K.; Washington, G. Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review. International Conference on Human-Computer Interaction. Springer, 2022, pp. 421–440.
4. Alabi, A. Introduction: Nollywood and the global south. *The Global South* **2013**, *7*, 1–10.
5. Umukoro, O.E.; Eluyela, F.; Inua, O.I.; Babajide, S. Nollywood accounting and financial performance: Evidence from Nigerian cinemas. *International Journal of Financial Research* **2020**, *11*, 271–280.
6. Ezepue, E.M. The new Nollywood: Professionalization or gentrification of cultural industry. *Sage Open* **2020**, *10*, 2158244020940994.
7. Danladi, S.S. Language policy: Nigeria and the role of English language in the 21st century. *European Scientific Journal* **2013**, *9*.
8. Benzeghiba, M.; De Mori, R.; Deroo, O.; Dupont, S.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; Rose, R.; others. Impact of variabilities on speech recognition. Proc. SPECOM, 2006, pp. 3–16.

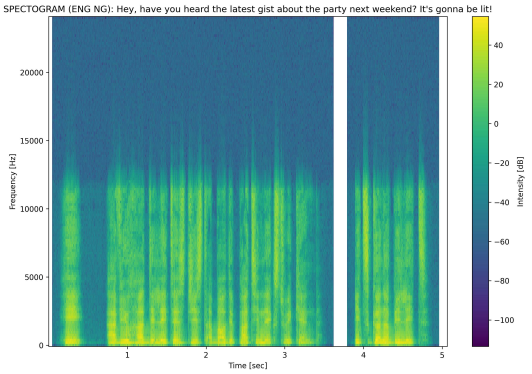
 doi:10.20944/preprints202402.0845.v1

9. Amuda, S.A.Y.; Boril, H.; Sangwan, A.; Ibiyemi, T.S.; Hansen, J.H.L. Engineering Analysis and Recognition of Nigerian English: An Insight into Low Resource Languages. *Transactions on Engineering and Computing Sciences* **2014**, *2*, 115–128. doi:10.14738/tmlai.24.334.
10. Babatunde, O. Automatic Speech Recognition for Nigerian-Accented English **2023**. doi:10.36227/techrxiv.24265738.v1.
11. Oluwatomiyin, S.; Misra, S.; Wejin, J.; Agrawal, A.; Oluranti, J. A Hybrid Translation Model for Pidgin English to English Language Translation. Data, Engineering and Applications; Sharma, S.; Peng, S.L.; Agrawal, J.; Shukla, R.K.; Le, D.N., Eds.; Springer Nature Singapore: Singapore, 2022; pp. 385–394.
12. Oladipupo, R.; Akinfenwa, E. Educated Nollywood artistes' accent as a Normative Standard of English pronunciation in Nigeria: Analysis of the phonemic realisation of educated Nollywood artistes. *English Today* **2023**, *39*, 207–217. doi:10.1017/S0266078422000207.
13. Karpf, A. *The human voice: How this extraordinary instrument reveals essential clues about who we are*; Bloomsbury Publishing USA, 2006.
14. Wunder, E.M.; Voormann, H.; Gut, U. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *icame Journal* **2010**, *34*, 78–88.
15. Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; others. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* **2024**.
16. Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Authur, R.; Bogin, B.; Chandu, K.; Dumas, J.; Elazar, Y.; others. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint arXiv:2402.00159* **2024**.
17. Barrault, L.; Chung, Y.A.; Meglioli, M.C.; Dale, D.; Dong, N.; Duquenne, P.A.; Elsahar, H.; Gong, H.; Heffernan, K.; Hoffman, J.; others. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596* **2023**.
18. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. International Conference on Machine Learning. PMLR, 2023, pp. 28492–28518.
19. Agarwal, M.; Agarwal, S.; Anastasopoulos, A.; Bentivogli, L.; Bojar, O.; Borg, C.; Carpuat, M.; Cattoni, R.; Cettolo, M.; Chen, M.; others. Findings of the IWSLT 2023 evaluation campaign. Association for Computational Linguistics, 2023.
20. Chen, W.; Yan, B.; Shi, J.; Peng, Y.; Maiti, S.; Watanabe, S. Improving massively multilingual asr with auxiliary ctc objectives. ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

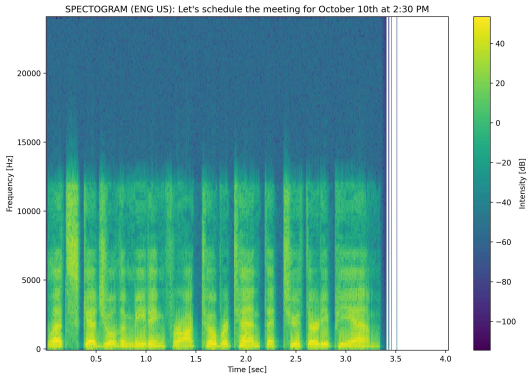
 doi:10.20944/preprints202402.0845.v1



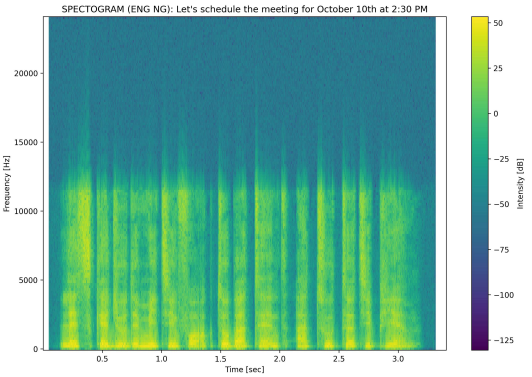
(a) Spectrogram for Sentence 1 (US English)



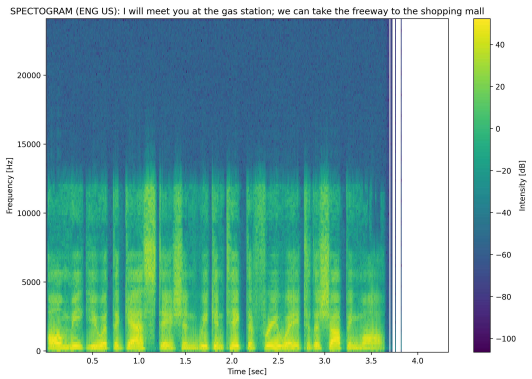
(b) Spectrogram for Sentence 1 (NG English)



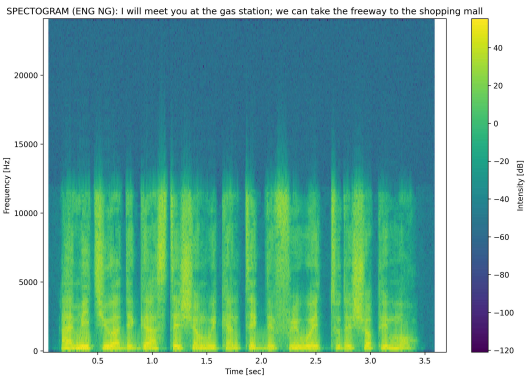
(c) Spectrogram for Sentence 2 (US English)



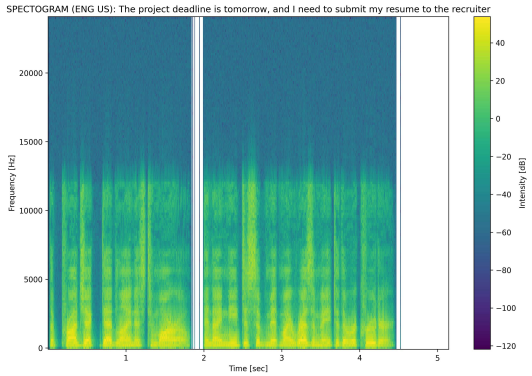
(d) Spectrogram for Sentence 2 (NG English)



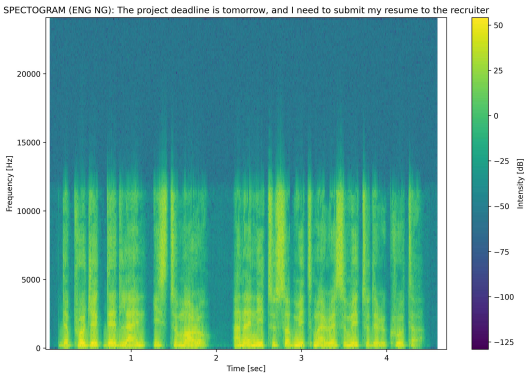
(e) Spectrogram for Sentence 3 (US English)



(f) Spectrogram for Sentence 3 (NG English)



(g) Spectrogram for Sentence 4 (US English)



(h) Spectrogram for Sentence 4 (NG English)

Figure 1. Spectrogram comparison of four sentences in English spoken by speakers from the USA and Nigeria.