Bias against English-Speaking Africans in Automated Speech Recognition

William Lu, Aditya Singh
wlu98761@usc.edu, apsingh@usc.edu
Department of Data Science, University of Southern California, Los Angeles, California 90007, USA
(Dated: April 30, 2024)

Automated Speech Recognition (ASR) systems have significantly improved human-computer interaction, enabling seamless use in applications like voice assistants and customer service. However, these systems often underperform for diverse linguistic groups, particularly English-speaking Africans, due to biases stemming from non-diverse training datasets. Our study examines these biases by evaluating the OpenAI Whisper-Small model, fine-tuned with the AfriSpeech-200 dataset to better recognize African-accented English. The results demonstrate the effectiveness of using targeted datasets to reduce bias, suggesting a path towards more equitable ASR technologies. This research contributes to discussions on fairness in AI and proposes actionable steps for developing inclusive speech recognition systems.

Keywords: Machine Learning, Automated Speech Recognition, Fairness, Bias

I. INTRODUCTION

Automated Speech Recognition (ASR) systems, with their wide range of applications across various fields, have significantly improved people's lives. From virtual assistants like Apple's Siri and Amazon's Alexa to voice commands on mobile phones and automated online job interviews that screen applicants with speech recognition, their impact is notable. However, ASR's effectiveness varies among different groups of people. Research has shown that ASR systems make more errors when interpreting words spoken by African Americans and other English-Speaking Africans compared to those spoken by white individuals. This bias in speech recognition can have a detrimental impact on African-English speakers, leading to doubts and concerns about identity, race, and fairness. Now we see there is bias against African spoken English in ASR, this issue can also happen to other minority groups or people who speak with non-native-English accents. Hence, it is important that more researchers should investigate and resolve this issue. [1]

Our research will focus on the bias existing in ASR, specifically the acoustic differences between English spoken by African Americans and other Africans compared to white individuals, and a potential solution to reduce this bias: fine-tuning OpenAI's Whisper ASR system. We will primarily focus on two datasets. The first is the LibriSpeech ASR Corpus, which contains 1000 hours of English speech mostly spoken by white people. The second is the AfriSpeech-200 dataset from HuggingFace, specifically focusing on English and South African English accents. By utilizing these datasets, we aim to delve into the nuanced speech patterns and vernacular specific to Afro-English speakers and explore potential methods to resolve this issue.

II. LITERATURE REVIEW

A. Dataset Overview

In 2020, Koenecke et al. found that automated speech recognition systems developed by Amazon, IBM, Google, Microsoft, and Apple made twice as many errors when interpreting English words spoken by African Americans compared to those spoken by white individuals. [1] They argued that these speech recognition systems exhibit racial bias.

Koenecke's research utilized two corpora of conversational speech: the Corpus of Regional African American Language (CORAAL), which consists of sociolinguistic interviews with dozens of Black individuals speaking African American Vernacular English (AAVE) to varying degrees, and Voices of California (VOC), another series of interviews. According to Koenecke's research team, audio data is challenging to collect, and sociolinguistic analysis often spans significant time gaps. The CORAAL dataset was recorded digitally from 2016 in Washington DC and Rochester, except for Princeville, which was recorded on cassette tape in 2004, while the VOC dataset was recorded from 2014 to 2017. Thus, the timing of the recordings and the recording media could contribute to bias in the results. Their analysis was limited to adult speakers with good audio quality, specifically excluding segments with background noise and ensuring snippets were of a manageable length, 5 to 50 seconds, for ASR systems. This resulted in a total of 4,445 snippets from Black speakers and 4,372 from white speakers. Additionally, to ensure consistency and facilitate error rate calculations, modifications were made to the ground-truth human transcripts, such as standardizing nonstandard spellings, removing unintelligible audio content flags, and standardizing the transcripts' formatting and spelling.

B. Method Overview and Analysis

Koenecke's research aimed to dissect racial bias in ASR systems through various methods. It measured bias using the word error rate (WER), defined as the total number of word substitutions, deletions, and insertions between machine and ground-truth transcriptions divided by the total number of words in the ground truth. The study first analyzed WER by segmenting it according to race. Black speakers had an average WER of 0.35, compared to 0.19 for white speakers. It then focused on the variation in error rates by location, given the nature of the CORAAL dataset. Data from black speakers in Princeville and Washington DC both had higher WERs than those from Sacramento and Humboldt County in the VOC dataset; however, Rochester, which consisted entirely of black speakers, had a WER comparable to the VOC. This led to an analysis of dialect density measures. Furthermore, when examining the language models of ASRs, snippets from black speakers showed lower perplexity, suggesting better predictability, which contradicts the higher WERs observed.

This discrepancy suggests that the issue may not lie with the language models (lexicon and grammar) but rather with the acoustic models of ASRs, which struggled with the pronunciation features unique to black speakers. Hence, based on this research, the bias in ASRs could stem from pronunciation and acoustic differences or accents between black and white speakers.

Research conducted by Shefali G et al in 2023. presents a method to reduce the WER disparity between African American English (AAE) and Mainstream American English (MAE) in ASR systems. The core of their methodology involves training an audio classifier to distinguish between AAE and non-AAE speech using a small amount of out-of-domain, long-form AAE data from CORAAL, YouTube, and Mozilla Common Voice. This classifier, built on a pre-trained speech foundation model and a 2-layer fully connected network, achieves high precision and recall in identifying AAE speech. It is then used to select AAE-specific utterances from a large corpus of untranscribed, short-form queries. This selected data, combined with coarse geographic information indicating regions with a high prevalence of AAE, is used for semi-supervised learning to fine-tune the ASR model. [2] From this research, we learn the method to improve ASR speech recognition through not only mitigating training data bias but also using an audio classifier to distinguish between different ways of speaking the same language, AAE versus MAE. Thus, careful selection of data can reduce bias.

R. Dorn's work presents an innovative approach to enhancing ASR systems for AAVE through the development of dialect-specific models. The study underscores the significant discrepancies in ASR performance between MAE and AAVE, attributing these differences to phonetic, syntactic, and lexical variations inherent to AAVE. Dorn's methodology involves the adaptation of existing ASR models with a focus on these dialectal characteristics, employing both acoustic and linguistic modifications to improve recognition accuracy. The results, as detailed in the proceedings, indicate a marked improvement in ASR performance for AAVE speakers, highlighting the potential for more inclusive and effective speech recognition technologies. This research not only contributes to the technical advancement of ASR systems but also emphasizes the importance of linguistic diversity in technology development[3]. Dorn's investigation into dialect-specific models for AAVE not only showcases the potential for tailored ASR systems but also addresses a critical gap in speech technology's inclusivity. The study meticulously outlines the process of adapting existing ASR frameworks to better accommodate AAVE's unique linguistic features. By incorporating dialect-aware training materials and fine-tuning model parameters to align with AAVE's phonological and syntactical patterns, Dorn demonstrates a significant reduction in error rates for AAVE speech recognition. This research not only sets a precedent for the customization of ASR systems to various English dialects but also champions the cause of linguistic equity in technological applications. The success of Dorn's models serves as a compelling argument for the necessity of dialect diversity in speech recognition technologies, proposing a future where digital inclusivity is paramount.

This paper serves as a resource for investigating acoustic bias within ASR systems, emphasizing the impact of dialect diversity on speech recognition accuracy. Acoustic bias occurs when the system's performance varies significantly due to differences in phonetic and phonological characteristics of dialects. In the case of AAVE, certain phonetic features such as vowel sounds, consonant cluster reductions, and intonation patterns might be underrepresented or misrepresented in the training data, leading to higher WER for AAVE speakers. By examining how these acoustic differences contribute to bias, researchers can develop more dialect-inclusive ASR models.

In their 2024 study, Hamel and Kani explore the multifaceted factors influencing the performance of machine learning models in the automatic recognition of AAVE. This research identifies key variables such as the quantity and quality of training data, the architectural differences in machine learning models, and the incorporation of dialect-specific features as critical determinants of ASR accuracy for AAVE. By conducting a series of experiments with various model configurations and training datasets, the authors demonstrate that enhancements in data representativeness and model adaptability lead to significant improvements in recognizing AAVE speech. Their findings suggest a path forward for the development of more robust and equitable ASR systems, capable of accurately serving diverse linguistic communities. This paper is particularly valuable for its comprehensive analysis of the interaction between data, model architecture, and dialectal nuances, offering insights that could inform future research and development in the field[4].

But the study of bias in ASR for English speakers does not end here. The research conducted by A. DiChristofano et al. titled "Global Performance Disparities Between English-Language Accents in Automatic Speech Recognition" has shown that ASR technologies might inadvertently create accessibility barriers for individuals with accents that deviate from the system's training data, primarily focused on mainstream accents like those from the United States. [5] The research suggests that speakers born in countries that are political allies of the United States, and those whose first language is English, tend to receive better ASR service. This bias indicates a significant challenge in ensuring equitable access to technology, potentially affecting speakers of African English varieties and their ability to use ASR-dependent services effectively.

C. Next Step

Therefore, it would be insightful if we study more about the sounding difference between Afro-English Varieties and mainstream English, understanding the traits of audios. Investigating how variations in the quantity and quality of training data affect recognition accuracy for minority dialects can illuminate the presence of data representativeness bias. By examining the relationship between data diversity and model performance, strategies can be developed to improve data collection and curation processes. This approach aims to ensure that training data covers a broader spectrum of linguistic diversity, ultimately leading to more equitable and effective ASR systems.

III. METHODOLOGY

A. Bias Reduction Idea

ASR systems nowadays are powerful and have brought a lot of benefits to people. Nevertheless, as we have previously mentioned in the literature review sections, ASR systems have downsides, such as possessing racial bias in audio recognition. Therefore, to create or improve ASR systems, incorporating more data of African American spoken English and Afro-English varieties could potentially help us mitigate the bias existed in these systems because English-Speaking Africans will then have more representations. Since building an advanced ASR system from scratch involves weeks or even months of training and data collection, fine-tuning one of the best ASR system in the market, namely OpenAI's Whisper-small, becomes the core focus of the project.

B. Data

OpenAI's Whisper-small is one of the best ASR systems in the market currently. It enhanced ChatGPT's power to another level, aiding people to perform question answering even through speeches.

According to OpenAI's documentations on Whisper-small, the dataset OpenAI used to train the model included LibriSpeech, CORAAL, Common Voice 5.1, WSJ, CHiME-6, and many more. However, there seem to be only one dataset, CORAAL, that is majorly about African American English audios. All the other five datasets may contain African American and Afro-English varieties audios, but they do not seem to be representative enough.

Thus, to bring more presence for more English-Speaking Africans, we decided to use AfriSpeech-200 from HuggingFace as an additional dataset to improve Whisper-small ASR system. The AfriSpeech-200 dataset is a dataset with 120 African accents from 13 countries and 2,463 unique African speakers. The goal is to raise awareness for and advance African English ASR research, especially for the clinical domain. For the South African English accent data, we have 114 audios for fine tuning and 5 for testing. For the English accent data, we have 106 audios for fine tuning and 46 for testing.

Another dataset we decide to apply is LibriSpeech. This is a dataset that is white audio dominated. We aim to use 1000 audios from this dataset to study characteristics of English spoken by white and use AfriSpeech-200, specifically South African English and English accents, to study traits of English spoken by Africans and for fine-tuning Whisper-small from OpenAI.

C. Data Preprocessing

To conduct data analysis on the LibriSpeech and AfriSpeech-200 datasets and fine-tuning the ASR model by OpenAI require intensive data preprocessing and management.

For fine-tuning, we need to ensure that the input size and format of our data match that of Whisper-small. Furthermore, we need to drop all other features of the audios and only keep the audios and their corresponding transcripts.

We created a Whisper feature extractor to perform three main operations. Since speech audios for computer devices expect finite arrays, we have to manipulate speech audios by sampling values at fixed time steps, called sampling rate, which is usually measured in Hertz. We used the same sampling rate of 16kHz as Whisper-small to ensure the correct audio speed. Second, we pad or truncate a batch of audio samples such that all samples have an input length of 30 seconds. In other words, if an audio is shorter than 30 seconds, we add zeros to the end of the one-dimensional array. If an audio is longer than 30 seconds, then it is truncated to 30 seconds. Finally,

our Whisper feature extractor converts the padded audio arrays to log-Mel spectrograms to mimic the human auditory range.

D. Data Exploration

For data exploration, we focused on three areas of the audios in these two datasets. We aim to study the duration, average pitch, and energies of these audio files through plotting the distribution and thus find some differences.

Duration: The length of time an audio file lasts, typically measured in seconds.

Energy: The intensity or loudness of sound within an audio file, often measured as power per unit area.

Pitch: The perceived frequency of sound waves, determining whether a sound is perceived as high or low in tone.

Durations Histogram:

- LibriSpeech: Concentration 0-5 seconds, peak at 5 seconds, right-skewed.
- AfriSpeech: Spread 5-10 seconds, slight left skew, preference for shorter files.

Energy Histogram:

- LibriSpeech: Peak around 0.0025, indicating similar loudness, tails off quickly.
- AfriSpeech: Wider range, peak around 0.04, suggesting more variation in loudness.

Pitch Histogram:

- LibriSpeech: Central peak 100-150 Hz, tails off towards higher pitches.
- AfriSpeech: Evenly distributed 170-250 Hz, broader distribution, potentially reflecting more diverse speakers/styles

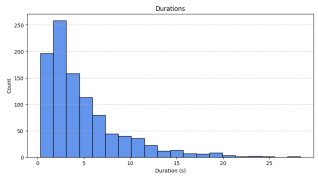

FIG. 1. LibriSpeech. Distribution of Duration.

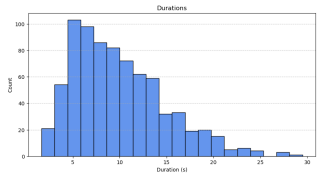

FIG. 2. AfriSpeech-200. Distribution of Duration.

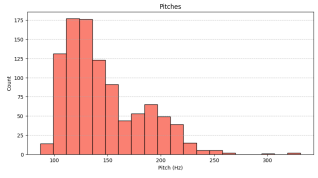

FIG. 3. LibriSpeech. Distribution of Average Pitch.

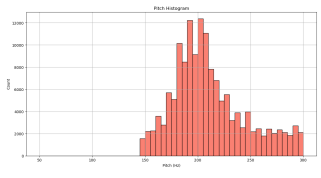

FIG. 4. AfriSpeech-200. Distribution of Average Pitch.

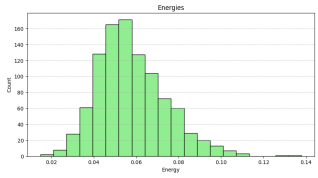

FIG. 5. LibriSpeech. Distribution of Energies.

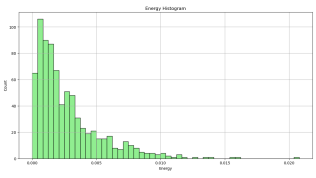

FIG. 6. AfriSpeech-200. Distribution of Energies.

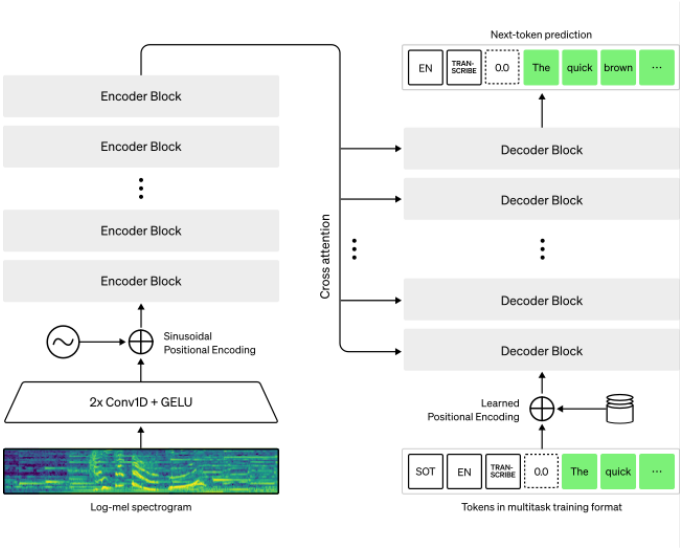

FIG. 7. Whisper Model Architecture. 6.

From the three sets of graphs, we can see that audios from the AfriSpeech-200 dataset tend to have more audios that are between five and ten seconds, whereas audios from LibriSpeech dataset are between zero and five seconds. For average pitch between the two datasets, audios from LibriSpech are more in the range of 100 to 150Hz. However, audios from AfriSpeech-200 are higher. Mostly are in the range of 170 to 250Hz. For energies, LibriSpeech dataset is almost normal but AfriSpeech is right skewed.

E. Model Building

After a thorough exploratory data analysis, we understand there are acoustic differences between English spoken by white and black people, mostly in terms of the energy and the pitch of their sounds.

Now, to create a model that can better recognize English spoken by African Americans and various African varieties, we have decided to use English and South African-accented English as the training data to fine-tune the Whisper-small model.

We chose Whisper-small because it is a lightweight version of larger Whisper models, making it efficient for us to use due to limited computational resources. It is robust to various types of noise and audio quality. As you

can see in Fig 7, Whisper-small adopts a sequence-to-sequence architecture, mapping sequences of audio spectrogram features into sequences of tokens. It uses the Whisper feature extractor to convert input audios into the log-Mel format. Then, several encoder blocks encode the log-Mel audios into hidden states. Following that are decoders, which autoregressively predict tokens, using cross-attention to complete the transcriptions.

First, we defined a data collector that can take preprocessed data from above and prepare PyTorch tensors for the model. Then, we continued to use the WER as the criterion to evaluate our model's performance. Remember, the WER is defined as follows, where S is the number of substitutions of words, D is the number of deletions of words, I is the number of insertions of words, and N is the number of words in a sentence.

$$\text{WER} = \frac{S + D + I}{N}$$

During the training process, we decided to experiment with different sets of parameters. For instance, we tested batch sizes of 8 and 16 for training, and per-device evaluation batch sizes of 4 and 8. We used learning rates of 2e-5 and 5e-6. Additionally, we set the maximum number of steps to 400 and the evaluation step to 25. We also enabled fp16 to utilize the GPU.

F. Results

For each experiment, since we are searching for the best parameters, it took an average of 3 to 5 hours each. The code for these experiments can be found by clicking here on GitHub.

TABLE I. Experiment Results in WER

	With Fine Tuning	Without Fine Tuning
South African English Accent Audios	23.65	25.53
English Accent Audios	53.28	49.13

In our first experiment, fine-tuning with South African English accent audios, we were able to achieve an improvement. Without adding more representative data to the model, we obtained a WER of 25.53. However, after adding the data to the model, we achieved a score of 23.65. Meanwhile, we observed a training loss around 0.15 and a validation loss of 0.66. In terms of losses, we may have overfit the model slightly; however, the improved performance still demonstrates the benefit of adding more underrepresented data to the model.

In our second experiment, fine-tuning with English accent audios, we did not observe an improvement. The addition of this set of audios resulted in a score of 53.28, whereas the original model scored 49.13. This result was surprising to us, as we expected it to work similarly to the first experiment. However, we observed both training and validation losses of about 0.11. We believe that the reason behind the lack of improvement in the second experiment was the inferior quality of the audio files, which made it difficult for the model to recognize the audios.

G. Conclusion & Future Work

Reducing bias in automated speech recognition is a task that requires efforts beyond computer science knowledge alone. It demands not only the collection of high-quality audio data but also data that are representative of diverse populations. While data augmentation is one method to reduce bias in ASR, increasing the availability of genuinely representative data for individuals of African descent is paramount. This lesson extends beyond English to encompass other languages as well. Fine-tuning a state-of-the-art ASR system can help mitigate racial bias between different groups, but this is just the beginning of achieving equity in the ASR system.

In future steps, fine-tuning Whisper-small with more data represents the most direct approach. Another approach is to develop an ASR system from the ground up using diverse and representative data from the beginning. In this way, we can ensure that the model being trained will be unbiased and fair. In addition, we can include training data of other minorities as well as these people likely face the same issue. This further shows the importance of data collection in building fair models.

Overall, bias in ASR systems is a serious issue in both current and future societies. Data scientists and engineers must remember and incorporate the concept of creating fair models and analyses to promote fewer misunderstandings, biases, and inequalities, fostering healthy and equitable relationships among people.

[1] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, Racial disparities in automated speech recognition, Proceedings of the National Academy of Sciences **117**, 7684 (2020), https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117.

[2] S. Garg, Z. Huo, K. C. Sim, S. Schwartz, M. Chua, A. Aksënova, T. Munkhdalai, L. King, D. Wright, Z. Mengesha, D. Hwang, T. Sainath, F. Beaufays, and P. M. Mengibar, Improving speech recognition for african american english with audio classification (2023), arXiv:2309.09996 [eess.AS].

[3] R. Dorn, Dialect-specific models for automatic speech recognition of African American Vernacular English, in *Proceedings of the Student Research Workshop Associated with RANLP 2019*, edited by V. Kovatchev, I. Temnikova, B. Šandrih, and I. Nikolova (INCOMA Ltd., Varna, Bulgaria, 2019) pp. 16–20.

[4] E. Hamel and N. Kani, Factors that influence automatic recognition of african-american vernacular english in machine-learning models, IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 509 (2024).

[5] A. DiChristofano, H. Shuster, S. Chandra, and N. Patwari, Global performance disparities between english-language accents in automatic speech recognition (2023), arXiv:2208.01157 [cs.CL].

[6] OpenAI, Whisper, `https://openai.com/research/whisper` (2023), accessed: your access date here.