INEQUITY IN POPULAR VOICE RECOGNITION SYSTEMS REGARDING AFRICAN ACCENTS

An Undergraduate Research Scholars Thesis

by

CHINAEMERE IKE

Submitted to the Undergraduate Research Scholars Program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Tracy Anne Hammond

May 2020

Major: Political Science

TABLE OF CONTENTS

ABSTRACT

Inequity In Popular Voice Recognition Systems Regarding African Accents

Chinaemere Ike
Department of Political Science
Texas A&M University

Research Advisor: Dr. Hammond
Department of Computer Science and Engineering
Texas A&M University

With new age speakers such as the Echo Dot and Google Home, everyone should have equal opportunity to use them. Yet, for many popular voice recognition systems, the only accents that have wide support are those from Europe, Latin America, and Asia. This can be frustrating for users who have dialects or accents which are poorly understood by common tools like Amazon's Alexa. As such devices become more like household appliances, researchers are becoming increasingly aware of bias and inequity in Speech Recognition, as well as other sub-fields of Artificial Intelligence. The addition of African accents can potentially diversify smart speaker customer bases worldwide. My research project can help developers include accents from the African diaspora as they build these systems. In this work, we measure recognition accuracy for under-represented dialects across a variety of speech recognition systems and analyze the results in terms of standard performance metrics. After collecting audio files from different voices across the African diaspora, we discuss key findings and generate guidelines for developing an implementation for current voice recognition systems that are more fair for all.

ACKNOWLEDGMENTS

I would like to thank Dr. Hammond, my amazing lab mates, especially Seth Polsley, and all those in the Sketch Recognition Lab for their help and support.

Thanks also go to my friends, especially Rebecca Schofield and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

I'd also like to thank these women: Joy Buolamwini, Kimberle Crenshaw, Megan Smith, Safiya Noble, Ruha Benjamin, Sasha Costanza-Chock and Rachael Tatman [1, 2]. These women are making powerful and impactful strides in the field of speech and voice recognition.

Finally, thanks to my family for all their encouragement. My dad is the reason why I wanted to pursue this topic in the first place and to him I'm grateful.

NOMENCLATURE

API	Application Programming Interface
ASR	Automatic Speech Recognition
EABC	Engineering Activities Building C
HCI	Human-Computer Interaction
TAMU	Texas A&M University
WER	Word Error Rate

CHAPTER I

INTRODUCTION

Motivation

In the age of up and coming technology, voice recognition systems are a hot commodity. Everywhere you go, there seems to be a commercial or a billboard related to "Alexa, what time is it?" or "Okay Google, can you tell me the temperature outside?" These devices are extremely convenient for the average person but only to an extent. The Google Home and Amazon's Echo Dot are both known for having issues picking up accents of non-native English speakers. The Washington Post's article "The Accent Gap" [3] noted that people with accents were getting left behind in the smart speaker phenomenon. Along with other papers and articles, research has demonstrated that not every voice is considered when developing voice recognition systems. A common theme we found in the literature was that accents from Africa were not tested. On a personal note from the author, what led to my own interest in this area of research started at home. My Nigerian father wanted to use my Echo Dot to set his alarm for work one day and realized Alexa could not understand him. I wanted to understand why that was the case and how I may be able to train Alexa to understand him. Furthermore, Joy Buolamwini at MIT Media Lab is doing similar work but with facial recognition. Gender Shades [4], her project where she discovered that image recognition systems did not pick up on darker shades, specifically black women, inspired this research project as well.

This led to the question: can more diverse voices be used to train voice recognition systems that can recognize African voices? We believe so. We hypothesize that performance can be improved by using audio samples from people from the African diaspora, as well as other non-native English speaking regions, to develop an implementation for current voice recognition systems. This project is important because it includes an entire demographic that has been partly neglected when creating voice recognition systems. It demonstrates the importance of diversity,

intersectionality [5], and inclusion in technology, a topic of extreme relevance as technology continues to be increasingly integrated into everyday life.

CHAPTER II

RELATED WORK

Accented English and Speech Recognition

Prior to conducting this research, we looked for similar work related to speech recognition and accented English. The first paper we found was “Using accent-specific pronunciation modelling for robust speech recognition" by Humphries, Woodland and Pearce [6]. Their research consisted of building a tree that was used to build a new pronunciation dictionary for use during the recognition process. The experiments they conducted presented for the recognition of Lancashire and Yorkshire accented speech using a recognizer trained on London and South East England speakers. The results showed that the addition of accent specific pronunciations can reduce the error rate by almost 20% for cross accent recognition. This is helpful information in relation to our work because of their trained recognizer.

Another paper in a similar realm is “A comparative analysis of UK and US English accents in recognition and synthesis" by Qin Yan and Saeed Vaseghi [7]. Their paper discussed a comparative study of the acoustic speech features of two major English accents: British English and American English. Their experiment examined the deterioration in speech recognition resulting from the mismatch between English accents of the input speech and the speech models. They performed a detailed study of the acoustic correlations of accent using intonation pattern and pitch characteristics. They realized that accent differences are acoustic manifestations of differences in duration, pitch and intonation pattern and of course the differences in phonetic transcriptions. Specifically, British speakers possess much steeper pitch rise and fall pattern and lower average pitch in most of vowels. Finally a possible means to convert English accents is suggested based on above analysis.

Next, we looked at papers that experimented with accents outside English descendants. Accent Detection and Speech Recognition for Shanghai-Accented Mandarin by Yanli Zheng,

Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Dan Jurafsky, Rebecca Starr, Su-Youn Yoon [8]. Their paper disussed a new approach that combines accent detection, accent discriminative acoustic features, acoustic adaptation and model selection for accented Chinese speech recognition. Ultimately, the results showed that their approach can improve the recognition of accented speech.

Similarly, a paper published at Columbia University surveyed a range of accents from different non-English speaking countries as well. In the paper, Automatic Dialect and Accent Recognition and its Application to Speech Recognition [9] Fadi Biadsy focused on automatically identifying the dialect or accent of a speaker given a sample of their speech, and demonstrates how such a technology can be employed to improve Automatic Speech Recognition (ASR). They described a variety of approaches that make use of multiple streams of information in the acoustic signal to build a system that recognizes the regional dialect and accent of a speaker. Initially, they analyzed the effectiveness of language identification that have been successfully employed by that community, applying them here to dialect identification. At the end of their research, they introduced several novel modeling approaches: Discriminative Phonotactics and kernel-based methods. They tested their best performing approach on four broad Arabic dialects, ten Arabic sub-dialects, American English vs. Indian English accents, American English Southern vs. Non-Southern, American dialects at the state level plus Canada, and three Portuguese dialects. They concluded that utilizing a linguistically-motivated pronunciation modeling approach can improve the Word Error Rate of a state-of-the art ASR system.

Industry Efforts

Fortunately, the tech industry and many researchers are discussing this topic of inequity in ASR systems and working towards quality improvements.

NPR

During the Fall of 2019, NPR began an experiment [10] asking people to lend their voices as their data. They want to know how well do ASR systems understand English speakers of all backgrounds.

Mozilla Common Voice

Common Voice [11] is Mozilla's initiative to help machines learn how the average person speaks. They understand that in order to build diverse voice systems, developers need an extremely large amount of diverse voice data. Their goal is to create a high quality publicly open data set with voices ranging from different backgrounds.

Voicing Erasure

Voicing Erasure [12] is a a poetic piece recited by champions of women's empowerment and leading scholars on race, gender, and technology. Led by computer scientist and digital activist Joy Buolamwini, Voicing Erasure highlights: 1. Voice Systems have biases [13], 2. Voices are being surveilled and not protected, 3. Voice Systems reinforce stereotypes and 4. Contributions in the field (by women) are being erased [14].

CHAPTER III

METHODOLOGY

The design and development of understanding Automated Speech Recognition system for my project will be conducted with a list of all transcriptions of the files from an open source audio file repository. We will compute accuracy across different dialects for some metrics. We combined all the Word Error Rate (WER) results into averages and statistically examined the differences by each dialect.

VoxForge Corpus and Different Dialects

The audio transcriptions we used came from VoxForge's corpus of files. VoxForge is an open source speech data set that holds a collection of transcribed speech from different languages and dialects from around the world. The audio files were recorded from the year 2007 to 2018. Each file contained 10 random sentences that were provided by VoxForge. All the files were listed as anonymous but the speaker had the option to include their gender. The files were recorded on VoxForge's Speech recorder system.

As mentioned above, there were thousands of languages ranging from English to Persian to select from. In the case of my research, we chose to use the English audio files. Over 6300 English files were available for use and within the files, each were classified by a specific dialect. The dialects we used were: American English, European English, British English, New Zealand English, South African English and Indian English. The South African dialect was the only dialect among the African diaspora. We decided to included Indian English as it was the only other non-European, labeled dialect that represented a significant portion of the corpus. VoxForge's data set was the only open-source data set of speech files that had a variety of different dialects. The initial goal was to collect speech files from different African diaspora languages but due to time constraints and resources available.

Parsing Audio Files

Next, once we decided which dialects to use, we downloaded each file and created a folder name after the origin of the dialect. Each dialect folder held over 80+ files. American English files were the most common and made up the majority of my corpus.

Using Python to read directory contents and process downloaded files, we were able to automatically generate transcriptions for each audio sample with the Speech Recognition module [15]. This module can be configured to operate with a multitude of free and commercial ASR systems, and it handles all of the networking, communication, and API calls in the background.

Selected ASR Systems

We investigated each of the available recognizers in the Speech Recognition package. Many commercial systems can become costly to use, although some have free credits available for students.

Sphinx

Originally named CMUSphinx, is a free open source transcription system for efficient speech recognition. CMUSphinx tools are designed specifically for low-resource platforms. Running the files through a non-commercial transcription system helps test its quality across many different applications.

WIT AI

Similar to Sphinx, Wit.AI is a free and open source transcription system that was acquired by Facebook in 2015. Wit.AI allows for developers to add a few lines of its code to instantly build in speech recognition and voice control.

Google Cloud

Among the several products Google offers, their Speech-to-Text transcription service was available for use and allotted users 300 USD in credit to transcribe files. It enables developers to convert audio to text by applying powerful neural network models in an easy-to-use API. The API recognizes more than 120 languages and dialects to support a global user base.

IBM

Lastly, IBM Watson's is a commercial cloud system that utilizes deep-learning AI algorithms to apply knowledge about grammar, language structure, and audio signal composition to create customizable speech recognition for optimal text transcription.

Others

There were other transcription systems we considered using but due to certain constraints, we did not use them.

Microsoft Azure

Microsoft Azure's Speech to Text service swiftly prevails common speech recognition barriers, such as unique vocabularies, speaking styles, or background noise while making audio more accessible by helping the world engage in conversations in real-time. The issue we faced here was gaining access to the API in order to run the transcription system on our end.

Houndify

Another service we used in the early stages was Houndify's Speech-to-Text system. We were able to run the files through the system with out any errors until more files were added to the corpus. Houndify became too restrictive on its allotted usage.

Performance Metrics

Once all the transcriptions for a given prompt were generated, we saved the results together with the original prompt text in a CSV file for easy comparison. The industry standard measure for ASR performance is Word Error Rate (WER). WER represents how many errors are present in a transcript by using Levenshtein distance, which compares the similarity of two sequences [16]. WER considers word-level differences to determine which words were correct, substituted, or dropped [17, 18].

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \qquad \text{(Eq. 1)}$$

for S substitutions, D deletions, I insertions, C correct words, and N total words.

The jiwer Python package [19] is used to compute WER for each result and save that score alongside the transcript in the final CSV output. The next section examines these results both qualitatively and quantitatively.

CHAPTER IV

RESULTS AND DISCUSSION

Using the methods described above, we generated a CSV file with transcriptions and WER scores for nearly 1500 prompts spread across a selection of ASR systems and English accents. From the averages and standard deviations across dialects and systems, there are several interesting results worthy of discussion.

Error Rate by Dialect

The key finding was the performance of each system on accented English versus the General American classification. Figure 1 shows the average WER scores grouped by dialect. It's important to note that shorter is better since WER is a measure of error. There is a little over 25% error on the American English dialects. European English and New Zealand English dialects are just about exactly at the half mark for error. South African English, British English and Indian English WER are all over the half mark. Figure 1 has colored these dialects in red. This means that, on average, South African accents are misunderstood by ASR systems more than half the time!

The British English poor performance was unexpected considering its origins are in Europe. We can assume the samples contained some very regional accents that were difficult for the transcription systems to understand. On the other hand, South Africa and Indian are located in different parts of the world where their native languages are not English. The Indian English had the highest WER and the poorest performance out of all the dialects. It was quite surprising to see but it also highlights the need to include and train non-English origin dialects in the data set. If consider relying on ASR systems for entertainment or even hands-free control while driving, such poor performance could lead to potentially frustrating scenarios but also dangerous ones.

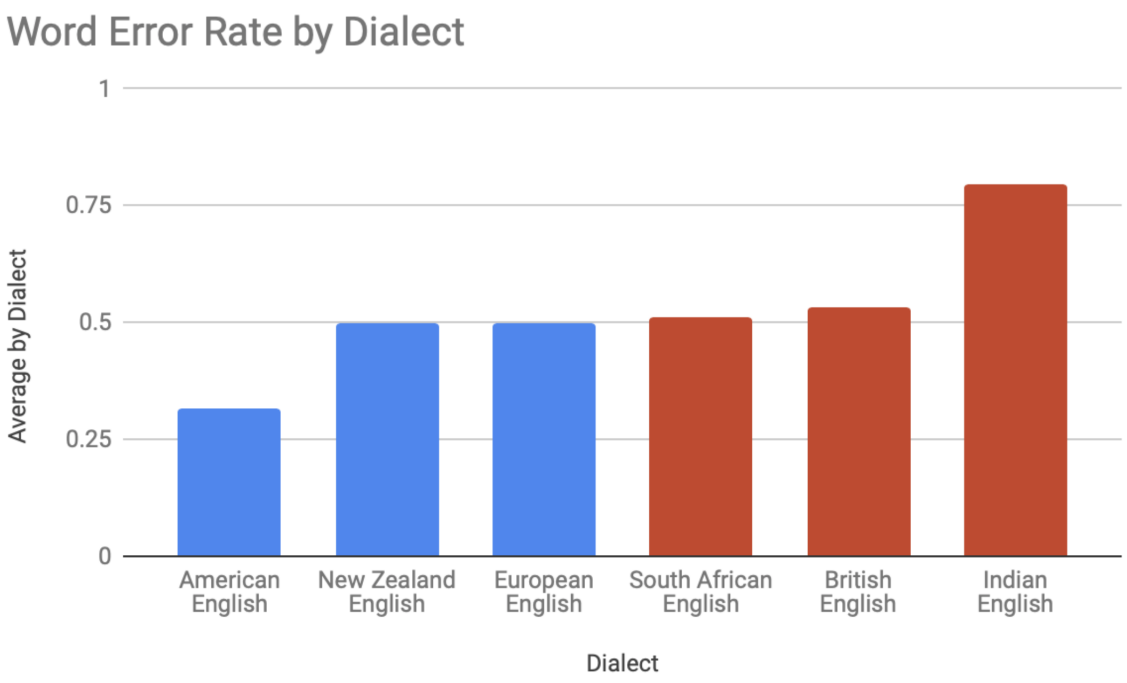

Figure 1: Word error rate by dialect, with those having over 50% error colored in red.

Error Rate by System

Another interesting visualization is the error by system across each dialect, as seen in Figure 2. Again, shorter is better. The reverse would be true for the standard measure of word accuracy, $1 - WER$.

On average, the commercial systems outperformed free ones by 17%. This is not a surprising result given the increased efforts companies expend towards improving their speech recognition systems, but it speaks to the value of optimizing and balancing data sets. For instance, IBM's cloud recognizer shows a notable drop in errors when recognizing European English versus other systems. This suggests that their training set includes more European accents than the other ASR systems. The goal moving forward would be to develop a transcription system that recognized more accents from Africa, starting with Nigeria.

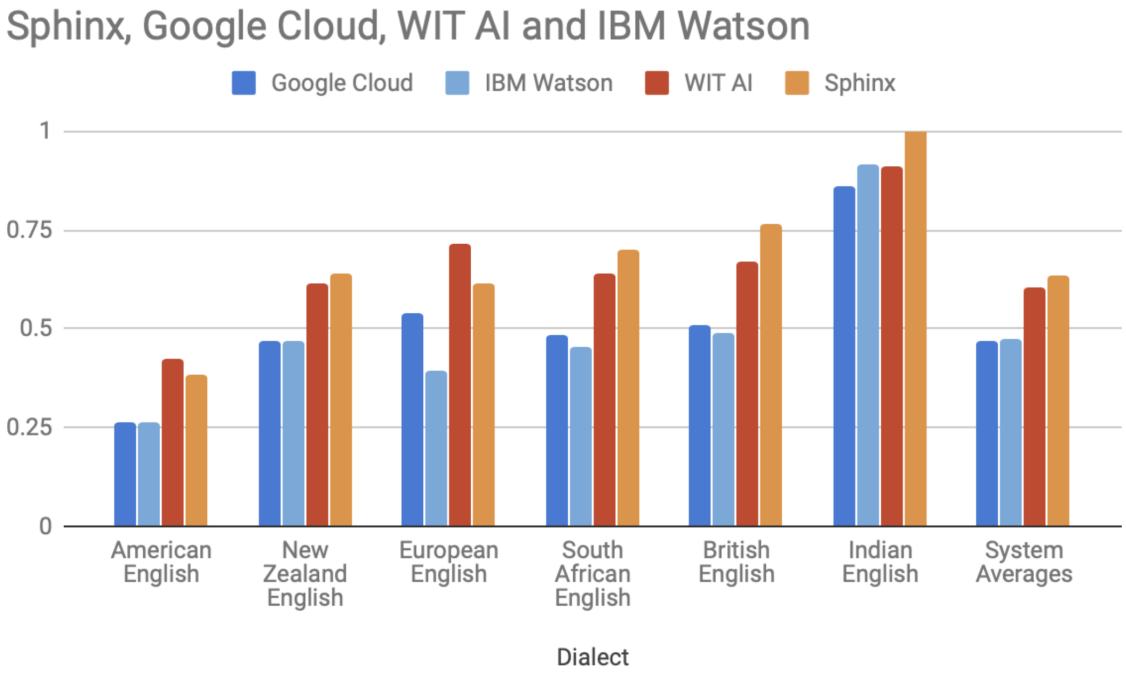

Figure 2: Word error rate by system.

Corpus Representation

Even in this study, the issue of lack of balance is readily apparent. We examined the overall contributions to the free VoxForge corpus and determined that nearly 50% of audio was labeled as General American English. The distribution is shown in Figure 3. The imbalance of different dialects such as this could be a source of potential bias if, for instance, an ASR included this corpus in training. This further motivates the need to collect more diverse and representative data for use in training machine learning systems.

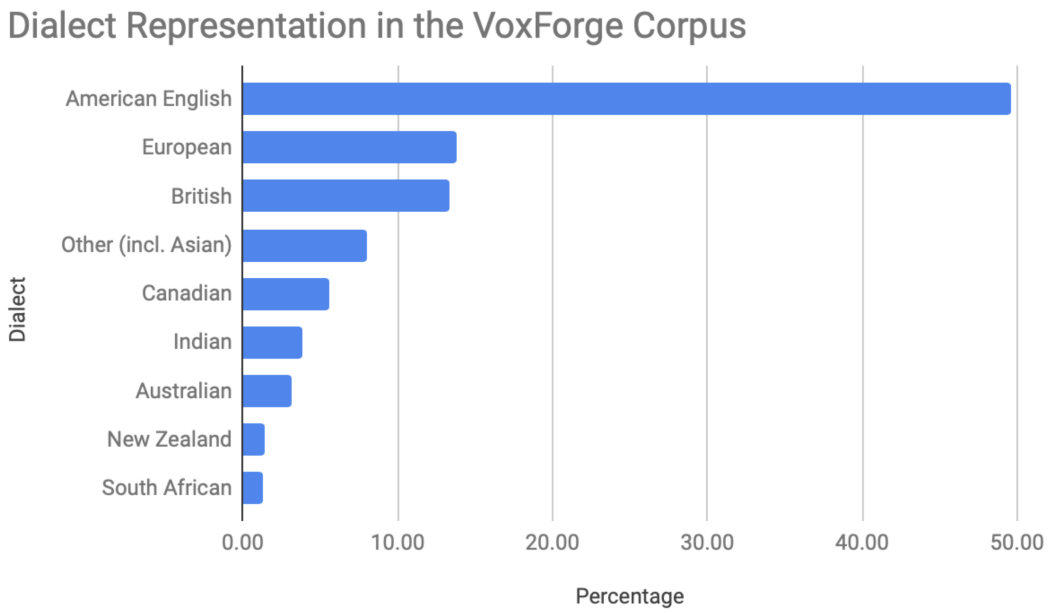

Figure 3: Dialect representation in VoxForge corpus, showing that nearly half of all data is for American English.

Other Findings

Speaker variation does play a role in ASR performance. We found standard deviation to be quite high across all systems, approximately 20%. In addition to making efforts to balance training data, further improvements to recognition algorithms may help reduce the impact of speaker variation.

A similar yet smaller practice study was conducted prior to this project where our friends and family of Nigerian descent (non-native English speakers) and had them recite several voice commands to Amazon's Echo Dot Smart Speaker. Alexa had a strenuous time comprehending certain commands especially the ones related to figuring out the weather forecast or the traffic in a certain city. It's quite discouraging considering these systems are being deployed for hands-free driving, phone support, smart home systems and other important things.

Overall, we need to collect more data from varying dialects and populations to improve performance. This will lead more equity and even better performance across different groups.

CHAPTER V

CONCLUSION

In this work, we realized that the current ASR systems in place far from inclusive. Throughout our results and discussion, we can see that American English speaker form the basis and the majority of transcription system data sets. Due to that fact, obtaining a balanced corpora can be tricky because even in the midst of this research, nearly half the data readily obtained is American English. This calls for the need of training more representative models.

The goal moving forward would be to develop speech recognition systems capable to either adapting well to other accents, or even normalizing all speech into some for of unaccented "computer language" that can be understood the same from everyone.

REFERENCES

[1] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 53–59, 2017.

[2] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions.," in *INTERSPEECH*, pp. 934–938, 2017.

[3] D. Harwell, "The accent gap," *The Washington Post*, Jul 2018. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

[5] D. W. Carbado, K. W. Crenshaw, V. M. Mays, and B. Tomlinson, "Intersectionality: Mapping the movements of a theory," *Du Bois review: social science research on race*, vol. 10, no. 2, pp. 303–312, 2013.

[6] J. Shearme and J. Holmes, "An experiment concerning the recognition of voices," *Language and Speech*, vol. 2, no. 3, pp. 123–131, 1959.

[7] Q. Yan and S. Vaseghi, "A comparative analysis of uk and us english accents in recognition and synthesis," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–413, IEEE, 2002.

[8] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for shanghai-accented mandarin," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[9] F. Biadsy, *Automatic dialect and accent recognition and its application to speech recognition*. PhD thesis, Columbia University, 2011.

[10] H. Jingnan, "Would you lend your voice to our experiment?," *NPR*, Dec 2019. https://www.npr.org/2019/12/25/784893309/would-you-lend-your-voice-to-our-experiment.

[11] "Common voice by mozilla." https://voice.mozilla.org/en.

[12] A. Koenecke, K. Crenshaw, M. Smith, S. Noble, R. Benjamin, and S. Costanza-Chock, "Voicing erasure," *AJLUnited*, 2020. https://www.ajlunited.org/voicing-erasure.

[13] 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*, 2020.

[14] C. Metz, "There is a racial divide in speech-recognition systems," *The New York Times*, Mar 2020. https://www.nytimes.com/2020/03/23/technology/speech-recognition-bias-apple-amazon-google.html.

[15] A. Zhang, "Uberi/speech_recognition," *Uberi*, Apr. 2020. https://github.com/Uberi/speech_recognition, original-date: 2014-04-23T04:53:54Z.

[16] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10(8), pp. 707–710, 1966.

[17] I. A. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," *IDIAP Tech Report*, 2004.

[18] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1–2, pp. 19–28, 2002. https://doi.org/10.1016/S0167-6393(01)00041-3.

[19] N. Vaessen, "jiwer: Approximate the WER of an ASR transcript," *jiwer*, 2020. https://github.com/jitsi/asr-wer/ access date: 2020-04-05.