Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Christina Annalice Chance April 12, 2022

Zoom Audio Transcription Accuracy for African American Vernacular English

By

Christina Annalice Chance

Dorian Arnold, Ph.D.
Advisor

Computer Science and Mathematics

Dorian Arnold, Ph.D.
Advisor

Emily Wall, Ph.D.
Committee Member

Talea Mayo, Ph.D.
Committee Member

2022

Zoom Audio Transcription Accuracy for African American Vernacular English

By

Christina Annalice Chance

Dorian Arnold, Ph.D.
Advisor

An abstract of
a thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelors of Science with Honors

Computer Science and Mathematics

2022

Abstract

Zoom Audio Transcription Accuracy for African American Vernacular English
By Christina Annalice Chance

As telecommunication is becoming a growing part of society, there is a concern for reliability and accuracy for all users. African American Vernacular English has been a dialect marginalized and forgotten by the Speech Recognition and Natural Language Processing community, thereby making most speech recognition tools less accurate for Black speakers. This study explores Zoom's closed captioning services for both African American Vernacular English and Standard American English to assess the accuracy amongst the different regional forms of AAVE as well as compare the overall accuracy between SAE and AAVE. Python's Asr Evaluation module was used to compute the edit distance. About 9 hours from both the CORAAL data-sets and Santa Barbra Corpus of Spoken American English we used; both data-sets possess conversational speech with linguistic sounds and stuttering. Results suggested that Zoom's closed captioning tool works more effectively for AAVE than for SAE based on the current data. To supplement that data in order to determine if the outcome of this work can be generalized to all closed captioning for video-conferencing tools, more formal speech samples were analyzed to assess the effect of outside compounding factors. The supplementary experiment showed contradicting results to the main study.

Zoom Audio Transcription Accuracy for African American Vernacular English

By

Christina Annalice Chance

Dorian Arnold, Ph.D.
Advisor

A thesis submitted to the Faculty of the Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements for the degree of
Bachelors of Science with Honors

of Computer Science and Mathematics

2022

Acknowledgments

I want to thank everyone for enduring this journey with me. It was long and rough, and I truly did not know if I was going to make it, but I did. I first would like to thank Dr.Arnold who took on becoming my advisor without hesitation even though I was doing work completely outside of his field. Thank you for trusting me, for pushing me, and for believing in the work I was doing and championing for me. Next, I would like to thank my committee, Dr. Wall, Dr. Mayo, Dr. Klein, Dr. Hue, and Dr. El-Sayed. You all helped shape my thought process, my approach, and my understanding of the field; this thesis would be tragic if not for you all. I lastly want to thank all of my friends and family that listened to me complain, cry, and stress over this thesis. Each of you has helped in ways that cannot be expressed and I truly appreciate the impact you have had during this journey.

Contents

1 Introduction **1**

2 Background & Related Work **3**

2.1 Related Works 3

2.2 Language Definitions 5

2.2.1 Standard American English 5

2.2.2 African American Vernacular English 6

3 Methodology **8**

3.1 Process 8

3.1.1 Evaluation Metrics 9

4 Experiment & Results **12**

4.1 Experimental Setup 12

4.1.1 Audio Data 12

4.1.2 Text Processing 14

4.1.3 Video-Conferencing Tool 14

4.2 Experiment 1 15

4.2.1 Analysis: Experiment 1 15

4.3 Experiment 2 17

4.3.1 Analysis: Experiment 2 17

4.4 Discussion 19

5 Conclusion **22**

5.1 Future Works 22

5.1.1 Continuation of Work 22

5.1.2 Fairness Space 24

Appendix A Full Data Mapping **25**

Appendix B Full Data Table **27**

Appendix C Full Supplementary Data **29**

Bibliography **31**

List of Figures

3.1 The generalized closed captioning transcription analysis process. . . . 9

4.1 Box plot of word recognition rate for regional AAVE representing: Atlanta, GA; Lower East Side Manhattan, New York; Washington, DC. 15

4.2 Box plot of word error rate for regional AAVE representing: Atlanta, GA; Lower East Side Manhattan, New York; Washington, DC. 16

4.3 The word error rate for all AAVE and SAE samples. 18

4.4 The word recognition rate for all AAVE and SAE samples. 19

List of Tables

2.1 Examples of a few common grammatical, syntactical, and phonological features of AAVE that differentiate from standard English sentence structure and linguistics sound [1]. Many of these features are reflected in other languages outside of English or have historic uses that remained in the dialect [2]. 6

4.1 The word error rate and word recognition rate for AAVE and SAE using the Asr_Evaluation module in python to calculate edit distance. 17

4.2 Word error rate, word recognition rate, and sentence error rate of the supplemental formal data for AAVE and SAE. 21

A.1 Audio file to reference text mapping filenames and audio length. . . . 26

B.1 Word error rate and word recognition rate for both AAVE and SAE samples, percentages in decimal form. 28

C.1 Audio length of supplementary data audio file. 30

Chapter 1

Introduction

Today, Machine Learning (ML) and Artificial Intelligence (AI) are becoming more omnipresent in the lives of everyday people. Tools and software that were originally used by specific demographics and classification of people are now becoming more commercial and open use. With this shift in target population arises a concern about the robustness and reliability of these tools, especially for under-represented populations. In many spaces, we have seen the effect of limited and non-diverse data contributing to and amplifying spaces of bias. Researchers like Joy Buolamwini and Timnit Gebru, in Gender Shades: Intersectional Accuracy Disparity in Commercial Gender Classification, highlighted the stark biases in the accuracy of gender classification tools against darker-skinned female individuals [3], or Ruha Benjamin who in her novel, Race After Technology: Abolitionist Tools for the New Jim Code, addresses the many sociological and cultural backings of many biased AI tools and their larger impact [4]. This work brings to light a few of the many instances where technology compounds biases.

Looking into the present-day as the current Covid -19 pandemic shifts our reliance more on technology than ever before, we are witnessing this problem propagate into how we communicate with one another. As we entered a time where in-person com-

munication is no longer the standard, telecommunication is growing in use and popularity. Large companies are shifting to online meetings, colleges are migrating to virtual classes, and families are connecting through video conferencing apps to stay connected during these distant times. As the users of video-conferencing tools are evolving, the accuracy and reliability of the tools should also evolve. Tools like Zoom and Google Meets, which were once used for business conferencing and professional meetings, are now being used in households around the world for people to stay connected. With that is a shift in the language and dialects being used on the platform. Larger platforms like Zoom added live captioning features recently and have been further developing the feature as it has become more of a necessity during the pandemic. However, we are witnessing performance issues for many marginalized and underrepresented groups. As the core of closed captioning and transcription services is rooted in Automated Speech Recognition and Natural Language Processing technology, we see the propagation of limited dialect and accent data for marginalized groups affects the overall accuracy and usefulness of these features.

These major concerns of accessibility and reliability for these closed captioning features on video-conferencing tools has led to this work. We want to assess the accuracy of Zoom's closed captioning features on African American Vernacular English compared to Standard American English as well as analyze the performance of this features on the various regional forms of African American Vernacular English. Utilizing Zoom's close captioning feature, screen share, and meeting recording, python's Asr_Evaluation module and JiWER module, the Corpus of African American Language, and the Santa Barbra Corpus of Spoken American English, we are able to collect hypothesis and reference text samples for both dialects and gather metrics to assess their overall accuracies. While work in the space of dialect-specific accuracy is not novel, the assessment of use cases in closed captioning has not been explored beyond YouTube.

Chapter 2

Background & Related Work

2.1 Related Works

Automated Speech Recognition (ASR) is an interdisciplinary sub-field of computer science and computational linguistics that utilizes machine learning to translate speech into text. It uses machine learning and artificial intelligence modeling techniques such as neural networks and hidden Markov models to model the phonetic and acoustic structure of words to predict the text of the speech samples fed into the system. The hidden Markov model creates a parametric model of the acoustic features of the text and the neural network completes the classification using the weights trained on the acoustic features [5].

Many models are performing at overall high accuracy for transcription tasks, however, the accuracy depends on the speaker and their dialect. These models tend to have poor performance for dialects outside of Standard American English, making the accessibility and use of these tools limited. As ASR tools are becoming more integrated into society with use cases in healthcare, virtual banking, telecommunication, and other widely used spaces, the concern for improved accuracy is becoming more prominent. Dialect-specific models have become the alternative in an attempt

to create models that support dialects other than SAE. However, these alternatives are not being integrated into commercial tools.

There are various existing works on analyzing the accuracy of speech-to-text models on African American Vernacular English. Many of these works highlight commercial-grade tools like Mozilla's Deep Speech and Google's Speech-to-Text with the goal of exposing statistical biases in the tools. Koenecke et. al found that the average word error rate for black speakers across various automated speech recognition tools was 0.35 while white speakers had a 0.19 word error rate. They suggest that this gap in accuracy can be attributed to the acoustic models for these tools having a lack of black speakers samples to train the model itself, therefore not allowing the acoustic model to pick up on the nuanced difference in the phonetic and phonological features of African American Vernacular English [6]. Similarly, Martin and Tang found that one distinctive grammatical feature of AAVE, the habitual "be", as well as its surrounding words are more error-prone than the non-habitual case suggesting a lack of instances of the feature in the training data-sets.

To address this issue, a commonly suggested alternative is the dialect-specific model. Dorn explains that applications that are purposed for diverse speech and have pre-gathered text data would benefit from a dialect-specific model as the increase in accuracy of the model would outweigh the increased processing time to identify the dialect [7]. Continued work in improving the accuracy of the processing time of mixed dialect-specific models has led to great developments. Yoo et. al tackles the issue of larger volumes of data necessary to support a multi-dialect model by creating an acoustic model that generalizes for all dialects by adapting the model based on dialect information and internal representation. Hirayama et. al created pronunciation dictionaries that map the use of different terms and words to dialects allowing the model to calculate the probability of the dialect of the sample based on occurrences and frequency of certain words in different dialects [8]. In a later work, Hirayama et.

al provided an alternative approach from the pronunciation dictionary that was more inexpensive. They suggest a model that uses statistics to formulate vocabulary transformation using transformation weights that represent the likelihood of that sample being in that dialect. These works were done with the goal of making future use cases of automatic speech recognition tools more reliable and accessible for non-SAE dialects [9].

With the shift to virtual communication due to Covid-19, the use of closed captioning and transaction services for video conferencing tools is now being assessed to ensure accuracy and standard performance for all users. Joe Zaghloul, a chief revenue officer for AsseblemyAI conducted a benchmark assessment of different speech-to-text APIs transcription services for various media types that utilized Zoom. The findings suggested that these APIs have higher accuracy for tutorials and more professional-based uses than for casual and less structured media where the accuracy was relatively poor [10]. This has great implications on who can use these tools and for what these tools can be used for. However little work had been done to address or even assess the accuracy of these closed captioning tools that are now supporting hundreds of thousands of people throughout the pandemic, leaving a gap in the statistical accuracy of these tools.

2.2 Language Definitions

2.2.1 Standard American English

Standard American English (SAE), also called Academic English or Mainstream English, is the baseline language that other English dialects are said to have been created from. As speakers of SAE are systemically empower, this is the language taught in academia and classrooms that dictates overall grammatical rules. However, SAE, despite being the standard in the United States, is normalized and used in predominately

white middle and upper class environment which discounts many other people who speak dialects like African American Vernacular English, Chicano English, Hawaiian Pidgin, and many other dialects specific to the United States. This standardization is then an erasure and invalidation of these other dialects as it deems them as improper and vulgar. [11]

2.2.2 African American Vernacular English

Linguistic Features	**SAE → AAVE Examples**
consonant clustering	West Side → "wes side"
"th" sound	this, that → dis, dat
habitual "be"	he is usually working → he be workin
exclusion of conjugated "be" verb	"where is she going" → "where she going"
negative concord	"I haven't seen anything" → "I ain't see nothing"
preterite "had"	"he went to work and then he called his client" → "he had went to work and then he had called his client"

Table 2.1: Examples of a few common grammatical, syntactical, and phonological features of AAVE that differentiate from standard English sentence structure and linguistics sound [1]. Many of these features are reflected in other languages outside of English or have historic uses that remained in the dialect [2].

African American Vernacular English (AAVE) is a dialect of the English language spoken by predominantly Black Americans throughout the United States. AAVE, also known as Ebonics or Black English, has many regional variations based on immigration patterns and ancestry but possesses similar grammatical and phoneme features. The dialect is also a toned language meaning that different inflections in speech sig-

nify different meanings. In Table 2.1, we see some of the most common grammatical, syntactical, and phonological feature of AAVE.

Chapter 3

Methodology

3.1 Process

In order to assess the accuracy of the various video-conferencing closed captioning tools against AAVE, we utilize the baseline of SAE to compare the accuracy. The procedural approach is a three-step process that entails data collection, pre-processing, and data analysis.

The data collection step includes the use of specified video-conferencing tools to produce transcripts. The method to access the closed captioning transcript varies based on platform, but the generalize process includes playing the audio file while closed captioning capability is enabled as well as recording the meeting. This will allow the feature to then transcribe the sample and produce a transcript of the captions throughout the recorded meeting.

This transcript possesses time-frames and unnecessary grammatical and syntactical features like punctuation and non-linguistic sounds. In order to make this data more comparable to the reference scripts provided by the speech corpus, we must clean both the closed captioning script as well as the reference script. Reference text files contain speaker names, timestamps, utterance indices, and other feature markers to

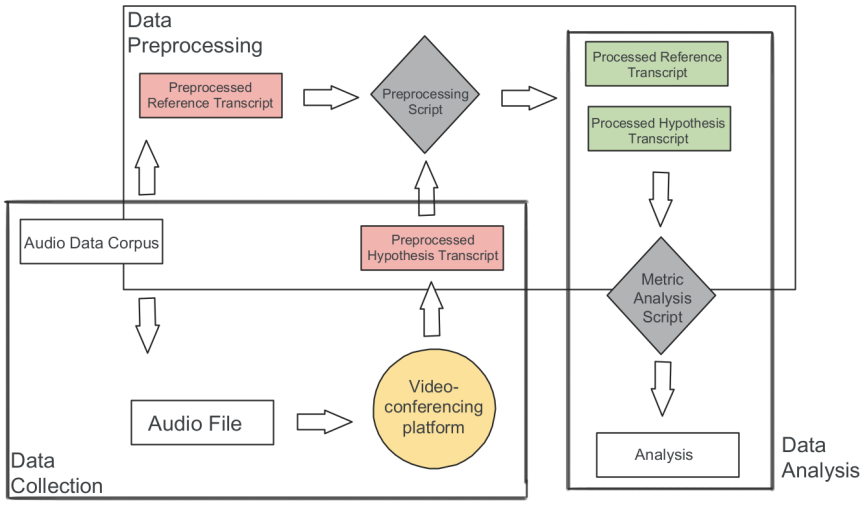

Figure 3.1: The generalized closed captioning transcription analysis process.

help identify when and who spoke. The pre-processing step sanitizes both the reference and hypothesized text files to reduce any non-linguistic differences using the script written with functionality support of python's JiWER module.

The data analysis step utilizes ASR_Evaluation to analyze the data. The ASR_Evaluation tool is utilized to extract the list of confusing words that include deleted, inserted, and substitutes terms as well as produce a summary of the word error rate and word recognition rate for each file.

3.1.1 Evaluation Metrics

For accuracy analysis, we are utilizing Asr_Evaluation, a python module used for automated speech recognition hypotheses[12]. This module uses edit distance (Levenshteinian distance) and is modeled after the align.c program used in the Sphinx ASR community. The main metric used in the program are word error rate, word recognition rate, and sentence error rate.

$$\textbf{Word Error Rate } (WER) = \frac{S+D+I}{N1} = \frac{S+D+I}{H+S+D} \tag{3.1}$$

where I is the total number of insertions, D is the total number of deletions, S is the total number of substitutions/replacements, N1 is the total number of reference words, and H is the total number of hits/successes.

The word error rate is a very common metric for analyzing the performance of speech recognition systems. The metric quantifies the percent difference between the reference sample and hypothesized samples using a similar approach as the Levenshtein distance. This metric does not have an upper bound as it is not deletion and insertion symmetric.

$$\textbf{Word Recognition Rate } (WRR) = \frac{H}{N1} \tag{3.2}$$

where N1 is the total number of reference words and H is the total number of hits/successes.

The word recognition rate, also known as word accuracy, is the complement of the word accuracy rate. The WRR is the percent of matched words between the reference and hypothesized sample. This metric has values ranging from 0.0 to 1.0.

$$\textbf{Sentence Error Rate } (SER) = \frac{S1}{S} \tag{3.3}$$

where S1 is the total number of incorrect statements and S is the total number of sentences.

The sentence error rate is the percent of sentences, or utterances, that were incorrectly transcribed.

Other metrics such as word information loss, relative information loss, match error rate and many others within Speech Recognition. The tool used for this work assessed accuracy on word error rate and word recognition rate which limited the scope in different approach of analysis. Also, with the goal of making the outcome of this

work as understandable and readable as possible, the current metrics used are more straightforward. Metrics like relative information loss not only require understanding of concepts like Shannon Entropy, but also has extreme complexities in calculation that make it unsuitable for live transcription use.

Chapter 4

Experiment & Results

In this section we will be elaborating on the experimental setup, the questions leading the experiment, and analyzing the outcome of the two experiments as well as overall descriptive statistics to better understand the data. We will explore the meaning of the results as well as comparing the outcome of the different experiments.

4.1 Experimental Setup

4.1.1 Audio Data

The data-sets utilized for the main study are the Corpus of Regional African American Language (CORAAL) and the Santa Barbara Corpus of Spoken American English. CORAAL is a data-set developed as part of the Online Resources for African American Language project (ORAAL) through the University of Oregon. The data-set currently contains regional accents of AAVE from Washington, DC; Princeville, North Carolina; Rochester, New York; Atlanta, Georgia; Detroit, Michigan; Lower East Side, New York; and Valdosta, Georgia. It is composed of over 150 socio-linguistic interviews from speakers born between 1891 and 2005 [13]. The Santa Barbara Corpus is a collection of natural conversations gathered across the United States with

the goal of capturing spontaneous speech. It has about 120,000 different words representing the recordings and a variety of identities in speakers. Both data-sets are conversational speech recognition training and testing data-sets [14]. Samples in both data-sets vary from 15 minutes to 1-hour long speech samples. Both also have similar quality including background noises and other features lowering the overall quality and clarity of the recordings. As the goal of this study is to assess the accuracy of the tool for everyday use for AAVE-speaking individuals, using conversational samples allows for a more realistic experimental setup. This work utilized about 9.5 hours of speech from each data-set with samples of an average length for the CORAAL data-set and Santa Barbara of about 37.5 minutes and 23.3 minutes respectively. The specific samples used for the study are specified in Table A

For this approach, samples will not be segmented into utterances but will be fed as a full sample. This reduces the likelihood of misplace utterance segmentation and handles the ill-structured transcription produced by the video-conferencing tools.

For the supplementary study the data-sets are Mozilla's Common Voice for more formal SAE samples and a self-produce sample set for more formal AAVE samples. Mozilla's Common Voice is a corpus of volunteer contributed utterances. The goals of this online open source and open contribution corpus is to "mobilis[e] people everywhere to share their voices." The data-set has about 14,000 validated hours of speech, 18,000 recorded hours of speech, and supports 87 different languages [15]. The self-produce sample set is shaped by three different speakers from different regional forms of AAVE and accents in the United States: Boston, Massachusetts; Kansas City, Missouri; and Hartford, CT. This supplementary work utilized only about 3.5 minutes of speech from each data-set with samples of an average length for the Common Voice data-set and self-produced of about 3 seconds and 1 minute respectively. The sample size is small as we wanted to explore the affects of the environment and use case of the data-set as it relates to accuracy amongst these two dialects.

4.1.2 Text Processing

As both the transcripts for the hypothesized and referenced samples are in 3 different formats, cleaning and centralizing the format and features of the samples is imperative to allow for a more fair comparison and analysis process. The reference transcripts and hypothesized transcripts are cleaned to remove punctuation, lowercase all letters, expand numbers to word form, expand contraction, and remove personal or redacted information, pauses, and non-linguistic sounds. To support the cleaning process, all text files are sanitized using the same methods. JiWER, a python module used to calculate similarity measures for automatic speech recognition evaluation, is used for the sanitation process[16]. The pre-processing functions provided by the module help address some of the major cleaning steps for text data like contract and numeric expansion and white space management.

4.1.3 Video-Conferencing Tool

For this work, we will be focusing on Zoom, version 5.9.1. Zoom is a video-telephony software program that allows for conferencing, webinars, meetings, lectures, and many other use cases. Zoom's live transcription and closed captioning services provide a variety of approaches ranging from assigning a member of the meeting to transcribe, using a third-party service, or utilizing Zoom's own live transcription services. In the most recent update in their documentation, they shared the limitations around their live transcription services. They shared that the tool has limitations including excessive background noise, clarity of speaker, and lexicons specific to geography. While Zoom provides these limitations for transparency, the statistics around these limitations do not exist, causing a gap in the qualitative understanding of these limitations. While many other video-conferencing services share the service provider their closed captioning process utilizes, whether it is Google Meets that uses Google's Text-to-Speech or Webex that uses Voicea's Enterprise Voice Assistant (EVA) speech

recognition technology, however, Zoom does not share their provider.

4.2 Experiment 1

Experiment 1: How accurate is Zoom's closed captioning services for different regional AAVE dialects?

For this question, we assessed the accuracy of the captioning services for three different regional forms of AAVE: Atlanta, GA; Washington, DC; Lower East Side Manhattan, New York. The three regions chosen had the most distinct regional differences and tonal accents from those present in the CORAAL dataset. Each region shares similar linguistics and syntactical features but includes regional-specific grammar and words. The goal was to overall evaluate the system for AAVE as well as explore how regional variations in AAVE can affect the overall performance of the systems.

4.2.1 Analysis: Experiment 1

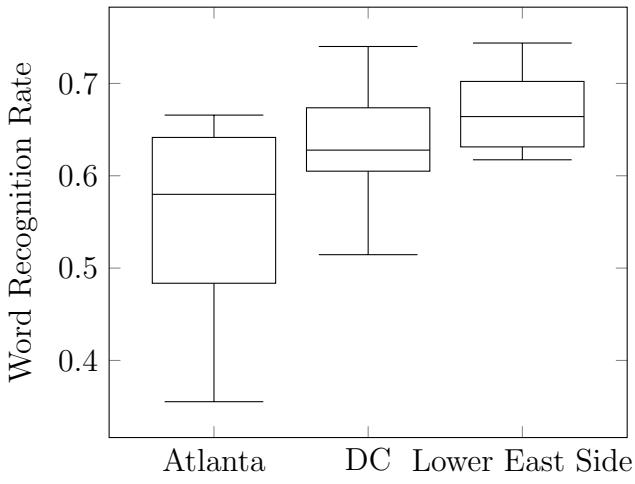

Figure 4.1: Box plot of word recognition rate for regional AAVE representing: Atlanta, GA; Lower East Side Manhattan, New York; Washington, DC.

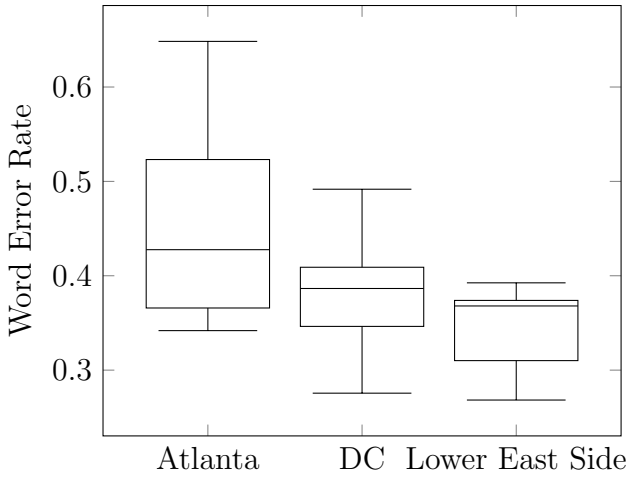

Figure 4.2: Box plot of word error rate for regional AAVE representing: Atlanta, GA; Lower East Side Manhattan, New York; Washington, DC.

Within the AAVE samples, the average word error rate for speakers from Atlanta, Georgia was 46.13% compared to Lower East Side Manhattan, New York whose word error rate is 34.255%, showing a significant difference between the accuracy amongst different regional accents with more northern regions like Lower East Side and DC having overall lower word error rates. We see in Figure 4.2 and Figure 4.1 a negative and positive correlation, respectively, between the range of the data and the means of each region. This suggest that with more variability in accuracy will alternatively decrease the accuracy for that regional form of AAVE. The results of this data suggest that certain components of the regional forms of AAVE have an influence on the overall accuracy. Using the confusion matrix produced by Asr_Evaluation to better understand the editing done by the algorithm for the text, more common words in AAVE are more prevalent to being edited more frequently in speakers from Atlanta than from DC or the Lower East Side.

4.3 Experiment 2

Experiment 2: How comparable are the accuracies of Zoom's closed captioning services amongst AAVE speakers and SAE speakers?

For this question, we assessed both the accuracy of the captioning service for SAE and AAVE and compared outputs to possibly witness overall performance differences between the dialects. Additionally, we wanted to assess the overall performance of conversational speech for the captioning service, as that has an impact on the quality and clarity of the audio.

4.3.1 Analysis: Experiment 2

	Average WER	Average WRR
AAVE	0.38986	0.62238
SAE	0.44081	0.57187

Table 4.1: The word error rate and word recognition rate for AAVE and SAE using the Asr_Evaluation module in python to calculate edit distance.

The results show that AAVE outperforms SAE in word recognition rate by 1.088 times and in word error rate SAE is more error prone by 1.13 times. We also witness the variance of SAE being 2.2 times larger AAVE suggesting that there is less variability and dispersion of error for the AAVE samples than there is for the SAE samples. The difference in the performance of the language between these two dialects was demonstrated in the type of editing done to the hypothesis text. For AAVE, many significant features of the dialect were edited out like the the various contractions of words like "wanna", "gonna", and "em". Additionally, words utilized frequently in the dialect like "bruh" were harder to identify in by the tool based off the confusion list produced for the samples. For SAE, some of the common edits done were, similar

to AAVE, some of the contractions like "cause" as well as tense changes for verbs. These results show that the tool overall does not perform well for conversational English as many of these edits produced are key features of colloquial speech. In any given samples there are about 100 deletions for "i"s alone and the insertion of articles each have counts of about 20. This shows that the tool can either be overly sensitive in which it is picking up sounds that are not speech or can be inadequately sensitivity in not being able to pick up common articles within an utterance. Additionally the variability in the spread of word error rates for both AAVE and SAE should be considered. In Fig 4.3, SAE has a larger dispersion of word error rate for their samples between their 1st and 3rd quarterlies compared to AAVE.

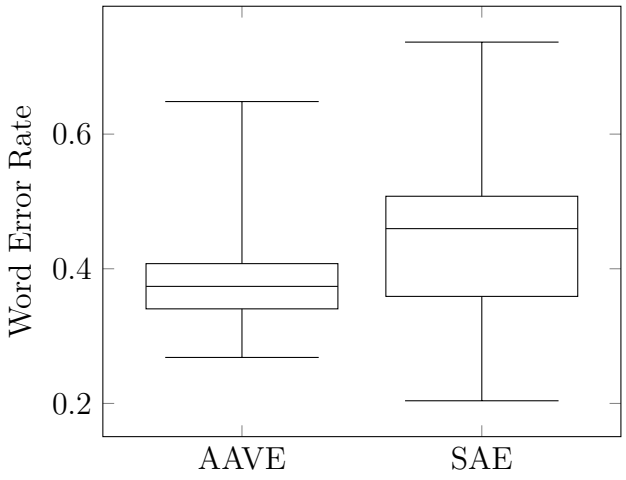

Figure 4.3: The word error rate for all AAVE and SAE samples.

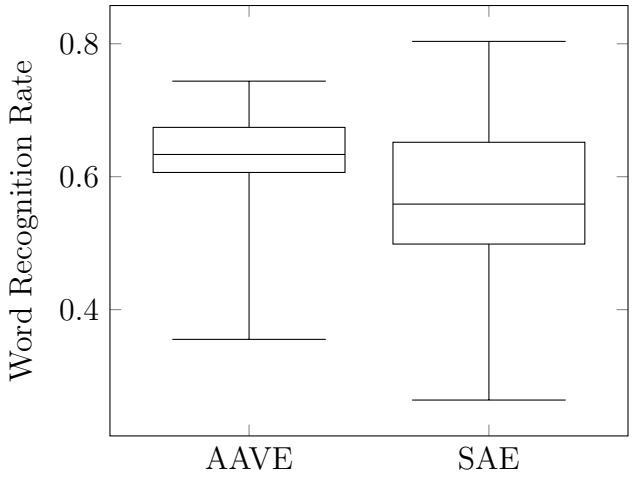

Figure 4.4: The word recognition rate for all AAVE and SAE samples.

4.4 Discussion

While the data suggests a significant difference between the performance of AAVE and SAE speakers with a favoring of AAVE speakers, which is not the expected outcome, there may have existed various compounding factors to led to such outcomes. Outside of the dialect being spoken, components of the audio file such as the ambient noise and linguistic noise, the varying conversational styles, and the quality of the recording itself may have contributed to the overall poor accuracy of Zoom. Additionally, the comparison between the two data-sets allows the question if they are as equatable in quality and type as originally assumed.

In order to better understand the comparability of the data-sets, we needed to understand the use cases and best way to categorize this form of speech audio. There is currently no standardized method or categorization rubric for grouping different types of speech data in the Speech Recognition and Natural Language Processing field as formal utterance based audio samples are considered the standard. However when considering the purpose of Zoom as well as the language of the newer users of the tool, everyday families and friends, it seemed necessary to stray from the more

formal business-like language originally utilized on this video-service. Therefore, a use of informal conversational audio samples like conversations during a family dinner or phone calls amongst friends seemed most ideal. However, as mentioned early when describing the data-sets, the CORAAL data can be considered more formal conversational samples as these samples were collected through an interview format with the goal of clearly documenting AAVE. Conversely, the Santa Barbara corpus can be considered as informal conversational data as it captures conversations between individuals in non-structured environments like in the car, in bed, or at the dining table. The difference in formal and informal audio collection can be a contributing factor to the levels or ambient noise as well as the clarity and volume of the speaker making it either easier or more difficult for Zoom to recognize someone speaking in the first place let alone trying to correctly distinguish what is being said.

Therefore, is it fair for us to use informal conversational data to assess the accuracy of Zoom's closed captioning system if it was not meant for informal speech? This is one of the major limitations around speech data, especially that of dialects. There is a lack of data for dialects like AAVE making it difficult to isolate different factors that contribute to the accuracy of the tools. The intention in choosing the Santa Barbra corpus was to a have a similar conversation type as that of the CORAAL data-sets and since CORAAL, although formally collected, possessed markers of more conversational speech like linguistic noises and everyday language, a comparable set is one that also possessed colloquial language. During the search for a comparable data-sets for CORAAL, the issue of structure in the ways of data collection and utterance as well as accessibility became an issue. As the standard for speech data in the field of Speech Reconnecting and Natural Language Processing is utterance based an formal, using data-sets purposed for linguistic studies, such as both the CORAAL and Santa Barbra corpus.

	Average WER	Average WRR	SER
AAVE	0.45867	0.62400	0.96154
SAE	0.07491	0.93258	0.56000

Table 4.2: Word error rate, word recognition rate, and sentence error rate of the supplemental formal data for AAVE and SAE.

With consideration of various contributing factors as to why accuracy overall was poor and additionally performance for SAE was worse than AAVE, it is necessary to frame what this outcome means. To reduce the variability in the quality of the data in order to focus on the variation in language as well as to be able to understand if this is a general case or if this outcome is explicitly valid for the conditions of this experiment. In order to assess if we are able to generalize these results as a standard for performance for Zoom's closed captioning service, we conducted a small supplemental analysis of more formal less conversational samples. These samples had no ambient noise and clear voice quality to reduce outside variables. In Table 4.2 the performance for each dialect compared to the main data is vastly different and shows a contradicting relationship compared to that discussed prior.

Chapter 5

Conclusion

In this work, we explored the accuracy of Zoom's closed caption tool for African American Vernacular English compared to Standard American English as well as accessed the accuracy of regional forms of African American Vernacular English. From the experiment results, we concluded that for the given data-set that Standard American English had a higher word error rate than African American Vernacular English and that the the regional form of AAVE for speakers from Atlanta had a higher overall word error rate compared to speakers from Lower East Side Manhattan and the Washington, DC. This means for the larger space that the dialect of the speaker affects the performance of the closed captioning service for Zoom which has larger implication around who and for what the tool can be used for.

5.1 Future Works

5.1.1 Continuation of Work

This work is only the beginning of what needs to be done specifically for assessing Zoom and for this project. The next steps include the gathering and use of comparable data for both AAVE and SAE speakers in order to neutralize all other

compounding features as well as to have intent pair samples to better understand the outcomes of the recognition and translation processes. Additionally, an aggregation of the confusion word list would allow a more intentional and qualitative analysis to provide more insight around what features of the dialect influence these edits and how the accuracy may be influenced by the acoustic and linguistic model for the tool. Further, an use of other metrics whether in from the Speech Recognition space or the space of artificial intelligence fairness would allow more robust analysis of the overall capability of the tool. AI Fairness 360, created by IBM released in 2018, is a open source toolkit of metrics used to assess bias and in data and models. It uses various metrics and algorithms from different researchers to centralize these tools with a goal of making them more accessible[17]. This toolkit can be essential in future works for finding ways to better utilize the power of algorithms to assess fairness.

In addition to the overall possible improvement of the analysis, another goal is to present more intentional counters to the AAVE samples. In future works, the overall accuracy of AAVE speakers can be assessed on a regional basis. With other dialects like Southern American English, a dialect spoken predominately by white people in the South, being a standard in their region, comparing the various regional AAVE dialects to their regions standard can help inform a better understanding the performance on a more fair basis. With that, also trying to understand the correlation between certain demographics of the region or city with the accuracy of speech. As seen in the the analysis of experimental question one, we witnessed that AAVE speakers from Lower East Side Manhattan performed better than those from Atlanta, Georgia. The question is could we see a clear correlation between the racial break brown of the city and the overall accuracy. Question like these may help in better formulating approaches to models that support different dialects.

5.1.2 Fairness Space

There remain several unanswered questions and concerns around this work, especially within assessing the overall accuracy of ethnic dialects on these large-use products like Zoom. Some next steps to consider are the lack of data for AAVE speakers and other ethnic dialects, the assessment of other video-conferencing tools marketed to the larger public, and addressing the models that support the speech recognition and transcription tools. These concerns and question can be challenged and answered in a myriad of ways including the creation of a corpus of formal US ethnic and cultural dialects, studying the effects of an AAVE trained acoustic model on the overall accuracy of the model, and assessing various use cases of video-conferencing as well as the different styles of speech they support. As the work in accessibility is endless, there is always something that can be done to improve another's experience.

Appendix A

Full Data Mapping

Original Audio/Text Name	Modified File	Audio Length (minutes)
ATL_se0_ag1_m_01_1	ATL1	45.56
ATL_se0_ag1_f_03_1	ATL2	30.09
ATL_se0_ag1_f_02_1	ATL3	36.52
ATL_se0_ag1_f_01_1	ATL4	31.03
DCB_se1_ag1_f_01_1	DCB1	35.06
DCB_se1_ag1_f_02_1	DCB2	37.02
DCB_se1_ag1_f_03_1	DCB3	48.46
DCB_se1_ag1_m_01_1	DCB4	45.55
DCB_se1_ag1_m_02_1	DCB5	65.53
DCB_se1_ag1_m_03_1	DCB6	42.12
LES_se0_ag2_f_01_1	LES1	26.29
LES_se0_ag2_f_01_2	LES2	27.27
LES_se0_ag2_f_02_1	LES3	37.1
LES_se0_ag2_f_02_2	LES4	0.59
LES_se0_ag2_m_01_1	LES5	49.06
SBC002 Lambada	SBC002	23.57
SBC003 Conceptual Pesticides	SBC003	26.07
SBC004 Raging Bureaucracy	SBC004	19.22
SBC009 Zero Equals Zero	SBC009	25
SBC012 American Democracy is Dying	SBC012	25.41
SBC015 Deadly Diseases	SBC015	26.06
SBC017 Wonderful Abstract Notions	SBC017	20.18
SBC019 Doesn't Work in this Household	SBC019	21.49
SBC023 Howard's End	SBC023	24.22
SBC028 Hey Cuties Pie	SBC028	25.17
SBC031 Tastes Very Special	SBC031	24.39
SBC033 Guilt	SBC033	10.47
SBC034 What Time is it Now?	SBC034	24.4
SBC035 Hold my Breath	SBC035	19.3
SBC036 Judmental on People	SBC036	26.51
SBC042 Stay out of it	SBC042	19.17
SBC043 Try a Couple Spoonfuls	SBC043	25.02
SBC044 He Knows	SBC044	29.07
SBC045 The Classic Hooker	SBC045	30.14
SBC047 On the Lot	SBC047	20.13
SBC050 Just Wanna Hang	SBC050	16.32
SBC058 Swingin' Kid	SBC058	25.47
SBC059 You Baked	SBC059	27.55
SBC060 Shaggy Dog Story	SBC060	24.5

Table A.1: Audio file to reference text mapping filenames and audio length.

Appendix B

Full Data Table

Samples	WER	WRR
ATL1	0.48142	0.52634
ATL2	0.34186	0.66570
ATL3	0.37384	0.63344
ATL4	0.64835	0.35530
DCB1	0.41039	0.60353
DCB2	0.40486	0.60925
DCB3	0.33907	0.68273
DCB4	0.36815	0.64626
DCB5	0.27558	0.74000
DCB6	0.49170	0.51452
LES1	0.39250	0.61729
LES2	0.31008	0.70217
LES3	0.26830	0.74375
LES4	0.36800	0.66409
LES5	0.37387	0.63126
SBC002	0.46100	0.54831
SBC003	0.49566	0.51040
SBC004	0.35762	0.65007
SBC009	0.35942	0.65706
SBC012	0.26698	0.74915
SBC015	0.40103	0.61173
SBC017	0.23860	0.77216
SBC019	0.43147	0.57762
SBC023	0.43732	0.56927
SBC028	0.23637	0.77013
SBC031	0.54641	0.45717
SBC033	0.69499	0.30695
SBC034	0.54388	0.46355
SBC035	0.62311	0.38236
SBC036	0.47935	0.52751
SBC042	0.73666	0.26402
SBC043	0.39277	0.62768
SBC044	0.20394	0.80356
SBC045	0.47106	0.63682
SBC047	0.23520	0.77624
SBC050	0.56813	0.43773
SBC058	0.47485	0.53305
SBC059	0.45822	0.54670
SBC060	0.46540	0.54558

Table B.1: Word error rate and word recognition rate for both AAVE and SAE samples, percentages in decimal form.

Appendix C

Full Supplementary Data

Sample Name	Audio Length (minutes)
common_voice_en_1	0.02
common_voice_en_2	0.02
common_voice_en_3	0.03
common_voice_en_4	0.02
common_voice_en_5	0.01
common_voice_en_6	0.02
common_voice_en_7	0.02
common_voice_en_8	0.04
common_voice_en_9	0.04
common_voice_en_10	0.02
common_voice_en_11	0.03
common_voice_en_15	0.04
common_voice_en_16	0.03
common_voice_en_17	0.03
common_voice_en_18	0.03
common_voice_en_19	0.05
common_voice_en_20	0.04
common_voice_en_21	0.03
common_voice_en_22	0.03
common_voice_en_23	0.04
common_voice_en_24	0.02
common_voice_en_25	0.03
common_voice_en_26	0.02
common_voice_en_27	0.03
common_voice_en_28	0.02
common_voice_en_29	0.05
AAVE_sample1	0.34
AAVE_sample2	0.4
AAVE_sample3	1.07

Table C.1: Audio length of supplementary data audio file.

Bibliography

[1] N. Ballister, "Linguistic features of aave."

[2] J. Rickford, *African American Vernacular English: Features, Evolution, Educational Implications*. Malden, Massachusetts & UK: Wiley-Blackwell.

[3] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification.," in *Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency*, 2018.

[4] R. Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*. USA: Polity, 2019.

[5] X. Tang, "Hybrid hidden markov model and artificial neural network for automatic speech recognition," in *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, pp. 682–685, 2009.

[6] 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*, vol. 117, no. 14, pp. 7684–7689, 2020.

[7] 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, (Varna, Bulgaria), pp. 16–20, INCOMA Ltd., Sept. 2019.

[8] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, "Automatic estimation of dialect mixing ratio for dialect speech recognition," in *Proc. Interspeech 2013*, pp. 1492–1496, 2013.

[9] N. Hirayama, S. Mori, and H. G. Okuno, "Statistical method of building dialect language models for asr systems," in *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pp. 1179–1194, 2012. Cited By :3.

[10] J. Zaghloul, "Comparing zoom transcription accuracy across speech-to-text apis," 2021.

[11] C. Chun, K. O'Neil, K. Young, and J. N. Christoph. University of Puget Sound.

[12] B. Lambert.

[13] T. Kendall and C. Farrington, "The corpus of regional african american language," 2021.

[14] D. Bois, J. W., W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson, and N. Martey, "Santa barbara corpus of spoken american english, parts 1-4," 2000-2005.

[15] "Mozilla common voice."

[16] *JiWER: Similarity measures for automatic speech recognition evaluation*, 2020 [Online].

[17] IBM, *AI Fairness 360*, 2018 [Online].

[18] S. Yoo, I. Song, and Y. Bengio, "A highly adaptive acoustic model for accurate multi-dialect speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5716–5720, 2019.

[19] D. Pallett, "A look at nist's benchmark asr tests: past, present, and future," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pp. 483–488, 2003.

[20] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 1, pp. 169–172 vol.1, 2001.

[21] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic speech recognition errors detection and correction: A review," *Procedia Computer Science*, vol. 128, pp. 32–37, 2018. 1st International Conference on Natural Language and Speech Processing.

[22] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, "Automatic speech recognition for mixed dialect utterances by mixing dialect language models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 373–382, 2015.

[23] A. Jørgensen, D. Hovy, and A. Søgaard, "Challenges of studying and processing dialects in social media," in *Proceedings of the Workshop on Noisy User-generated Text*, (Beijing, China), pp. 9–18, Association for Computational Linguistics, July 2015.

[24] N. F. Chen, S. W. Tam, W. Shen, and J. P. Campbell, "Characterizing phonetic transformations and acoustic differences across english dialects," *IEEE/ACM*

Transactions on Audio, Speech, and Language Processing, vol. 22, no. 1, pp. 110–124, 2014.

[25] Y. Halpern, K. Hall, V. Schogol, M. Riley, B. Roark, G. Skobeltsyn, and M. Baeuml, "Contextual prediction models for speech recognition," in *Proceedings of Interspeech 2016*, 2016.

[26] S. Tanberk, V. Dağlı, and M. K. Gürkan, "Deep learning for videoconferencing: A brief examination of speech to text and speech synthesis," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pp. 506–511, 2021.

[27] M. Lehr, K. Gorman, and I. Shafran, "Discriminative pronunciation modeling for dialectal speech recognition," in *Proc. Interspeech*, 2014.

[28] R. Tatman and C. Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," in *Proc. Interspeech 2017*, pp. 934–938, 2017.

[29] S. Kafle and M. Huenerfauth, "Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, ACM, oct 2017.

[30] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. Interspeech 2004*, pp. 2765–2768, 2004.

[31] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, (Valencia, Spain), pp. 53–59, Association for Computational Linguistics, Apr. 2017.

[32] S. Groenwold, L. Ou, A. Parekh, S. Honnavalli, S. Levy, D. Mirza, and W. Y. Wang, "Investigating African-American Vernacular English in transformer-based text generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 5877–5883, Association for Computational Linguistics, Nov. 2020.

[33] S. K. Pulipaka, C. K. Kasaraneni, V. N. Sandeep Vemulapalli, and S. S. Mourya Kosaraju, "Machine translation of english videos to indian regional languages using open innovation," in *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pp. 1–7, 2019.

[34] M. Pucher, N. Kerschhofer-Puhalo, and D. Schabus, "Phone set selection for HMM-based dialect speech synthesis," in *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, (Edinburgh, Scotland), pp. 65–69, Association for Computational Linguistics, July 2011.

[35] S. L. Blodgett and B. O'Connor, "Racial disparity in natural language processing: A case study of social media african-american english," 2017.

[36] H. Elfardy and M. Diab, "Sentence level dialect identification in Arabic," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Sofia, Bulgaria), pp. 456–461, Association for Computational Linguistics, Aug. 2013.

[37] S. McCarty, L. Pham, A. Thresher, A. Wasgatt, and E. Whamond, "Sorry, i didn't quite get that: The misidentification of aave by voice recognition software," 2021.

[38] C.-H. Lee, "Speech and audio processing for multimedia communications," in *TENCON '97 Brisbane - Australia. Proceedings of IEEE TENCON '97. IEEE*

Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (Cat. No.97CH36162), vol. 2, pp. 625 vol.2–, 1997.

[39] D. cheng Lyu, R. yuan Lyu, Y. chin Chiang, and C. nan Hsu, "Speech recognition on code-switching among the chinese dialects," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, 2006.

[40] J. L. Martin, "Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual 'be'," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 284, Association for Computing Machinery, 2021.

[41] S. Cho, F. Dernoncourt, T. Ganter, T. Bui, N. Lipka, W. Chang, H. Jin, J. Brandt, H. Foroosh, and F. Liu, "Streamhover: Livestream transcript summarization and annotation," 2021.

[42] E. Beldon, M. Tota, N. Williams, C. Vogler, and R. Kushalnagar, "Teleconference captioning accessibility,"

[43] A. Mbogho and M. Katz, "The impact of accents on automatic recognition of south african english speech: A preliminary investigation," SAICSIT '10, (New York, NY, USA), p. 187–192, Association for Computing Machinery, 2010.

[44] J. L. Martin and K. Tang, "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"," in *Proc. Interspeech 2020*, pp. 626–630, 2020.

[45] T. Jones, "What is aave?," 2014.

[46] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, "'i don't

think these devices are very culturally sensitive.'—impact of automated speech recognition errors on african americans,"