RACE AND REGIONALITY ON THE ASPIRE ASR MODEL

by

LILLIAN LE

(Under the Direction of MARGARET E. L. RENWICK)

ABSTRACT

Automatic speech recognition (ASR) enables the transcription of spoken speech into a written format. Previous research has shown racial biases in modern ASR systems exist and negatively affect Black speakers. In this thesis, speech data from the CallHome and CORAAL ATL, DCB, PRV, and ROC corpora are processed and given to ASpIRE, a DNN-HMM model built on the open-source ASR toolkit Kaldi. The trends in the model's word error rates between different phonological phenomena and corpora are considered in the context of the model's original training process and modern sociolinguistic knowledge. All in all, the training set used to develop the ASpIRE model is insufficiently enriched with phonological and lexical representations of AAL and Southern characteristics.

INDEX WORDS: Automatic Speech Recognition, African American Language, Computational Linguistics

RACE AND REGIONALITY ON THE ASPIRE ASR MODEL

by

LILLIAN LE

B.A., University of Georgia, 2020

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2021

RACE AND REGIONALITY ON THE ASPIRE ASR MODEL

by

LILLIAN LE

Major Professor:	Margaret E. L. Renwick
Committee:	John Hale
	Khaled Rasheed

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2021

DEDICATION

This one’s for me.

ACKNOWLEDGEMENTS

I would like to extend my deepest thanks to Dr. Renwick for her guidance and encouragement at every step of the way. This thesis would not exist without her expertise and empathy. I would also like to thank Dr. Hale and Dr. Rasheed for serving on my thesis committee.

Thank you, as well, to my friends and family who have cheered for me throughout this process.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background

African American Language (AAL) is arguably the most studied dialect of American English, but previous knowledge on the dialect stands to be expanded and explored in real-world applications, like automatic speech recognition (ASR). Now, AAL is a broad term meant to represent all of the varieties of English used by Black people in America. It encompasses African American English, African American Standard English, African American Vernacular English, Black English, Ebonics and so on. In general, African American Language can be distinguished by a distinct set of systematic grammatical and phonological characteristics, such as copula absence, invariant be, or final consonant cluster reduction (Rickford, 1999).

Previously, AAL has been thought of as uniform nationwide regardless of other regional influences (Labov, 1972; Wolfram and Fasold, 1974; Labov et al., 2006). However, this belief is under challenge and the availability of a publicly accessible, regional corpus of African American Language does much to further the investigation of regionality in African American speech (Wolfram and Kohn, 2015). The corpus referenced in the previous statement is the Corpus of Regional African American Language (CORAAL) which collects a multitude of smaller Black conversational speech corpora from cities around the United States (Kendall and Farrington, 2020).

ASR is the technology which enables the recognition and transcription of speech into a written format. It is increasingly being used for mobile voice assistants, automatic dictation and

transcription, handsfree automotive controls, and a variety of other tasks by millions of people. It is an immensely useful technology, especially for individuals experiencing decreased motor control in many cases. As its use becomes omnipresent, it is imperative that the ASR systems available are broadly inclusive and able to sufficiently serve different demographics in the population.

Previous investigations have shown that speaker characteristics affect ASR performance (Tatman, 2017; Tatman and Kasten, 2017; Koenecke et. al, 2020). Specifically, Koenecke et. al (2020) compares five ASR systems—developed by Amazon, Apple, Google, IBM, and Microsoft—with data from white and Black speakers across five US cities. Black speakers' utterances suffered higher rates of error under each system. Koenecke et. al (2020) stipulated some regional linguistic variation may account for the differences in ASR performance, but the association was not as clear. Nonetheless, regional dialect is certainly another facet of individual speaker characteristics that affects ASR performance. Tatman (2017) evaluates YouTube's automatic caption service across two genders and five dialects of English. Tatman and Kasten (2017) include a racial dimension alongside gender and dialect in a comparison of the Bing Search API and YouTube's automatic captioning. Both projects found a robust difference in accuracy across dialect. Additionally, Tatman and Kasten (2017) also saw higher error rates for non-White speakers.

As such, the general goal of this thesis is to use a modern ASR system to assess how race and regionality affect the model's performance and errors. The model selected is the ASpIRE model which operates on the open-source Kaldi toolkit (Povey et al., 2011) and was trained with the Fisher English corpus (Cieri et al., 2004). The data examined are subcomponents of CORAAL including speakers from Rochester, NY (ROC), Washington D.C. (DCB), Atlanta, GA (ATL), and

Princeville, NC (PRV). The performance of ASR on CORAAL is compared to the CallHome English corpus (Canavan et al., 1997), which serves as a representation of standard (White) American English speech. Over the course of the thesis, several research questions are considered and answered. (1) Does race affect ASR performance? (2) What specific features of AAL do ASR systems struggle with? (3) Do regional phonological patterns affect ASR performance? (4) Why does an ASR system perform differently on one type of speaker versus another?

1.2 Experiments and Results

To answer these research questions, the data is given to the ASpIRE model to process and transcribe through the Kaldi toolkit. The model extracts the Mel Frequency Cepstral Coefficients (MFCCs) along with other features of the audio data to produce an utterance labeled transcript of its most likely hypothesis for each utterance. Model performance is evaluated via Word Error Rate (WER) measurements on subsets of the data. WER is a measure of the number of full word matches between the hypothesized model transcript and the ground-truth human transcript. Formally, it is calculated as the summation of substituted, deleted, and inserted words divided by the total number of words in the ground-truth transcript. After scoring the hypothesized transcripts, we find the model performs the best on the CallHome data (WER = 23.99), then ROC (WER = 27.96), DCB (WER = 36.99), ATL (WER = 42.36), and lastly, PRV (WER = 50.53). A pairwise t-test with Bonferroni correction across the datasets showed that all pairwise WER averages were significantly different from one another ($p < 0.05$) except in the case of DCB and ATL.

Of the words scored, a subset of only monosyllabic words appearing in each dataset a minimum of ten times is taken for further investigation. Some common AAL and regional phonological features are selected to examine the specific effect they may have on ASR performance. Specifically, words with word-final consonant clusters ending in [t] or [d], the

voiceless dental fricative [θ], the voiced dental fricative [ð], postvocalic [l], or word-final postvocalic [r] are chosen to examine the effect of AAL on model performance. For regional variation analysis, words containing the vowels AY (/ai/, *ride*), EY (/eɪ/, *bait*), IY (/i/, *bee*), AH (/ʌ/, *bus*), AO (/ɔ/, *thought*), or AA (/ɑ/, *bot*) are selected.

Table 1. Overall and subset WERs

Dataset	Overall WER	Common Monosyllabic WER	AAL Words WER	Regionally Variable Vowels WER
CallHome	23.99	18.84	18.52	17.66
ROC	27.96	24.35	25.72	21.00
DCB	36.99	34.51	35.48	32.78
ATL	42.36	35.67	38.64	33.68
PRV	50.53	48.13	51.60	45.05

WER decreases from all words to only common monosyllabic words (CallHome = 18.84, ROC = 24.35, DCB = 34.51, ATL = 35.67, PRV = 48.13), as shown in Table 1. From this new baseline, WER increases in the words containing the phonemes associated with AAL for the CORAAL datasets, but not the CallHome dataset. On the other hand, WER improves overall for words containing vowels associated with different regional vowel shifts, but a steep downward trend persists as the corpus speakers become more Southern. These trends are directly tied to the ASpIRE model's training data. PRV is most poorly recognized because it features strongly Southern and Black speakers. DCB and ATL see less than optimal performance due to their mixed variety of dialects and transitional realizations of vowels as well. Ultimately, all of the CORAAL subcopora suffer from poorer model performance due to the speakers' African American identities and subsequent engagement with AAL.

1.3 Contributions

The discrepancies in WERs across the CORAAL and CallHome corpora found in the experiments demonstrate there is evidence of an underlying bias in the ASR model examined. The contribution of this thesis is the examination of the WERs produced by the ASpIRE model through a systematic isolation of specific phonological contexts where AAL and regional features are most likely. The trends in WERs between different phonological phenomena and corpora are considered in the context of the model's original training process and modern sociolinguistic knowledge, such as how speakers from different cities produce diphthongs. Additionally, lexical sources of error are considered within the same context, explicitly revealing gaps in the model's representation of Black and Southern speech. All in all, the training set used to develop the ASR model is insufficiently enriched with phonological and lexical representations of AAL and Southern characteristics.

1.4 Outline of the Thesis

In the following parts of this thesis, Chapter 2 introduces more background information, including descriptions on AAL, different regional dialects in the United States, the CORAAL corpus, the CallHome corpus, the fundamentals of automatic speech recognition and the toolkit Kaldi. Chapter 3 discusses the preprocessing and set up necessary to run the ASpIRE model. The model decoding and scoring parameters are explained as well. Chapter 4 lists the results in the form of word error rates per dataset. Chapter 5 takes a subset of the words found in the ground truth transcripts and compares their collective WERs against the inclusion of certain phonemes associated with AAL or regional dialects. Results in these comparisons are used to make inferences on the effect of a speaker's race and regionality on the performance of an ASR system. Chapter 6

provides a conclusion in an overview of the trends examined, their implications, and recommendations for future work.

CHAPTER 2

BACKGROUND

2.1 Characteristics of African American Language (AAL)

The term AAL has been defined as "an intentionally broad term meant to encompass all varieties of language use in African American communities" (Farrington 2020b). African American English (AAE) is a synonymous term that will be used interchangeably for this paper following the recognition that "most speakers of the variety see themselves, first and foremost, as speakers of English" (Britt and Weldon 2015). AAL differs from other varieties of English in two main dimensions: the sound (phonological) system and the grammar (morphosyntactic) system (McLarty 2020). There is an overlap of features between AAL and other varieties of English, such as Standard American English (SAE), white Southern American English, and Chicano English; however, AAL employs a unique combination of these features (McLarty, 2020).

A primary focus of this thesis project is the phonological system of AAL. Rickford (1999) and the ORAAL page "AAL Linguistic Patterns"[1] both provide detailed descriptions of grammatical features of AAL. Also based upon these sources and Lehr et al. (2014), Table 2 lists common phonological markers of AAL for consonants. The phonological rules are with respect to SAE, where the left-hand side[2] is the expected SAE realization. The usage of many of these rules can be sensitive to phonological contexts and speaker-specific social factors (Lehr et. al, 2014).

---

1 https://oraal.uoregon.edu/AAL/Linguistic-Patterns, cited as Farrington (2020a)

2 The phoneme preceding the → symbol

Table 2. Common phonological features of African American Language with linguistic rules mapping from Standard American English to possible African American Language realizations and examples

Phonological Description	Phonological Rule	Examples
Reduction of word-final consonants clusters, especially those ending in [t] or [d]	C → ∅ / C _ #	hand → han' desk → des' post → pos'
Devoicing of word-final voiced stops after a vowel	[-cont, +voice] → [-cont, -voice] / V _ #	pig → pik
Velar nasal fronting, where [ng] becomes [n]	[ŋ] → [n] / [ɪ] _ #	walking → walkin'
Variation of dental fricatives [θ] and [ð] as [t, f] or [d,v]	[θ] → [t] or [θ] → [f] [ð] → [d] or [ð] → [v]	thin → tin | bath → baf then → den | brother → bruvver
Reduction of /θr/ sequences as /θ/, especially before [u] or [o]	[r] → ∅ / [θ] _ {[u], [o]}	throwdown → thodown
Deletion or vocalization of /l/ after vowel	[l] → ∅ / V _ [l] → ə / V _	help → he'p toll → toah
Deletion or vocalization of /r/ following a vowel or between two vowels, especially in word-final position	[r] → ∅ / V _ {#, V} [r] → ə / V _ {#, V}	sister → sistuh for → fouh
Deletion of unstressed initial and medial syllables	σ[-stress] → ∅ / (σ) _ σ	afraid → 'fraid secretary → sec't'ry
Metathesis or transposition of adjacent consonants	V C1 C2 V → V C2 C1 V	ask → aks wasp → waps
Voiced fricatives (/v/ and /z/) as voiced stops (/b/ and /d/), especially in word-medial position before nasal	[v] → [b] / (V) _ (V, [+nasal]) [z] → [d] / (V) _ (V, [+nasal])	seven → seben isn't → idn'
Syllable-initial /str/ as [skr], especially before high front vowels	[t] → [k] / $ [s] _ [r] (V[+high, +front])	street → skreet destroy → deskroy

Neutralization/merger of [ɪ] and [ɛ] before nasals	[ɛ] → [ɪ] / _ [+nasal]	pen → pin
Lowering and backing of [i] to [æ] before velar consonants	[i] → [æ] / _ {[ŋ], [ŋk]}	thing → thang drink → drank

In addition to these primarily consonantal features of AAL, there exists the African American Vowel Shift (AAVS) (Thomas 2007; Kohn 2013). The AAVS most prominently involves the fronting of AA (/ɑ/, bot), and the raising and fronting of AE (/æ/, trap), EH (/ɛ/, bet), and IH (/ɪ/, bit) (Renwick and Olsen, 2017). Furthermore, monophthongal pronunciations of AY (/aɪ/, ride) and OY (/ɔɪ/, choice) have been documented (Rickford, 1999; Thomas, 2007; Renwick and Olsen, 2017). Southern African Americans may "exhibit a less diphthongal" AW (/aʊ/, plow) (Renwick and Olsen, 2017).

2.2 Characteristics of Regional American Dialects

There has been extensive research on the multiple varieties of American English (for overviews see Clopper et. al, 2005; Clopper and Pisoni, 2006; Labov et. al, 2006). The primary dialects represented in the data used for this project are from the Inland North, Southern, and Midland regions. Thus, this section will focus primarily on this selection.

Firstly, a vowel chain shift is a series of related sound changes. The movement of one vowel forces other phonemes in the vowel space to move so that the individual phonemes retain their distinctiveness (Chapman, 2017).

The Inland North dialect is characterized most typically by the Northern Cities Shift (NCS). The NCS is a vowel chain shift said to begin with the fronting and raising of AE (/æ/, *cat*). According to Labov (2010), the NCS subsequently sees

- the fronting of AA (/ɑ/, *bot*),

- the lowering and fronting of AO (/ɔ/, *thought*),
- the backing and lowering of EH (/ɛ/, *bet*),
- the backing of AH (/ʌ/, *bus*),
- and the backing and lowering of IH (/ɪ/, *bit*).

The NCS is primarily present around the Great Lakes region, spanning from Upstate New York to as far west as the "Twin Cities", Minneapolis and Saint Paul, Minnesota (Chapman, 2017). This area and its surrounding parts also present rhoticity, the pronunciation of /r/ as rhotic vowels or in syllable-final position (e.g., *car*).

Southern speech has only become studied somewhat recently compared to other American dialects. Labov began work on vowel quality in Southern White dialects in 1972 and proposed the "Southern Shift" or Southern Vowel Shift (SVS) as a system of vowel mutations found across the US South (Labov et al, 1972; Labov, 1991). The SVS begins with the monophthongization of AY (/aɪ/, *ride*) (Labov et al., 2006). Following this glide weakening, Labov et al. (2006) states the main elements of the SVS are

- the backing and lowering of EY (/eɪ/, *bait*),
- the rising and fronting of EH (/ɛ/, *bet*),
- the backing and lowering of IY (/i/, *bee*),
- the rising and fronting of IH (/ɪ/, *bit*),
- and the fronting (with some non-extreme raising) of AE (/æ/, *cat*).

Furthermore, EH and IH may develop prominent inglides such that "*sit* will be heard as equivalent to *see it* in Northern and Midland dialects and *set* as equivalent to *say it*" (Labov et al., 2006). The Southern Vowel Shift has been documented within the Southeastern United States from

Texas to West Virginia, although not always uniformly (Feagin, 2003; Labov et al., 2006). Namely, Atlanta has been commonly recognized as a divergent southern city with a complex vowel system that incorporates certain elements of multiple regional dialects (Labov et al., 2006; Kretzschmar, 2015).

Unlike the Northern and Southern dialects in the United States, the Midland dialect is characterized not by any prominent chain shifts but the lack thereof. Nonetheless, Labov et al. (2006) do list some general characteristics of Midland speakers:

- the fronting of OW (/oʊ/, *code*),
- the fronting of AW (/aʊ/, *mouth*),
- the neither completely present nor completely absent low-back "cot-caught" merger,
- and the raising and tensing of AE (/æ/, *cat*) before nasals.

Beyond vowels, the Midland is also firmly rhotic. The boundaries of the Midland dialectal region are not conclusively defined; however, it is widely accepted to reach from Ohio to central Nebraska and Oklahoma (Labov et al., 2006).

2.3 CORAAL Corpus

CORAAL is the first public corpus of AAL data and one of the few publicly available large-scale sociolinguistic data sets. It is a long-term corpus building project conceived in several smaller corpora, otherwise known as components. As of September 2021, CORAAL contains seven components featuring interviews from six different cities: Atlanta, GA, Washington, D.C., Lower East Side, NYC, Princeville, NC, Rochester, NY, and Valdosta, GA. There are two components which represent Washington, D.C, and with the exception of one of the D.C.

components, all of the interviews included in CORAAL took place between 2004 and 2018. The excluded D.C. component originates from Ralph Fasold's 1972 foundational study on African American Language in the D.C. area. Likewise, almost all of the components were created as part of separate projects, either dissertation research or local dialect studies. The 2016 Washington, D.C. subset was created specifically for CORAAL, but fills the 4 x 3 demographic matrix done for the 1968 Washington, D.C. subset.

CORAAL's transcription practices were adapted from the Sociolinguistic Archive and Analysis Project (SLAAP) (Kendall 2007). As such, its "transcripts align text to speech at a per-utterance level, where utterances are defined as uninterrupted speech sounds by the same individual, with utterances delimited at pauses" (Kendall and Farrington, 2020). Additionally, morphosyntactic variants were transcribed, but phonological variants were not. All CORAAL recordings have been anonymized. The CORAAL audio files themselves are generally 16-bit, 44.1 kHz, mono in WAV format.

At the beginning of this thesis project, the Valdosta, GA and Lower East Side, NYC components were not yet available. As such, they were not included in any experiments nor was the older Washington, D.C. component. Thus, the data utilized were the Atlanta (ATL), 2016 Washington, D.C. (DCB), Princeville (PRV), and Rochester (ROC) components. Interviewers in each of these datasets were also included in experiments. Some brief information on each component follows.

ATL consists of 13 primary speakers and one interviewer who make up a modern Atlanta friendship sphere (Farrington et. al, 2020). Many speakers were not born and raised in Atlanta but moved to the south from other areas. DCB is the largest of all of the CORAAL components included and consists of 48 primary speakers (Kendall et. al, 2018). Speakers were collected

through a friend of a friend network. All speakers were interviewed by Minnie Quartney. PRV consists of 16 primary speakers collected by Ryan Rowe, Walt Wolfram, and colleagues for the North Carolina Language and Life Project (Rowe, 2015; Rowe et al., 2018). Lastly, ROC consists of 14 primary speakers collected by Sharese King (King, 2018; King et al., 2020).

All four of these cities have significant African American populations and are historically significant in different ways. Atlanta was an important railroad and military supply hub during the Civil War. Today, its ever-growing population and status as a southern economic hub make it one of the South's most prominent cities. Like the speakers in ATL, many Atlanta residents are migrants from other parts of the country. Washington, D.C. has been the nation's capital since 1791. Its population is larger than that of both Wyoming and Vermont. Similar to Atlanta, Washington, D.C. sees many transplants from other areas; a little over a third of Washington, D.C. inhabitants are native to the city (U.S. Census Bureau, 2019). On the other hand, Princeville, North Carolina is the oldest town incorporated by African Americans in the United States and many of its present inhabitants are direct descendants of the town's founders (Farrington, 2021). Princeville is the smallest town in CORAAL by far and also has the highest percentage of African Americans in its population. Lastly, the Erie Canal runs directly past Rochester, New York whose growth and exposure to outsiders was greatly attributed to the canal's trade traffic (Erie Canal, 2021). A table with population and linguistic information on each of the cities is found below (Table 3).

Table 3. Total population, African American population percentage, geographic region, and expected dialect for Atlanta, GA, Washington D.C., Princeville, NC, and Rochester, NY

City	Total Population[3]	% African American[3]	Geographic Region	Expected Dialect[4]
Atlanta	498,715	47.22%	South	Transitional
Washington D.C.	689,545	41.45%	Mid-Atlantic	Transitional
Princeville	2,154	92.82%	South	Southern
Rochester	211,328	40.10%	North	Inland North

Despite its relatively new age, CORAAL has already been used in a moderate collection of studies. *American Speech* dedicated a special issue to the CORAAL components DCA and DCB in February 2019. That issue features seven investigations exploring grammatical variation, specific phoneme realizations, variable question intonation, identities in AAL (Kendall, 2019; Farrington and Schilling, 2019; Cukor-Avila and Balcazar, 2019; Grieser, 2019; Forrest and Wolfram, 2019; McLarty et al., 2019; Holliday, 2019; Quartey and Schilling, 2019). Farrington (2018) uses CORAAL to examine the relationships between vowel duration, final glottal stop replacement, and deletion of word-final [t, d]. Additionally, some components of CORAAL were used as the representation of Black speech in Koenecke et al. (2020) for a comparison of ASR performance across race.

2.4 CallHome Corpus

The CallHome English corpus of telephone speech was collected and transcribed by the Linguistic Data Consortium (LDC) primarily in support of the project on Large Vocabulary

3 Numbers as reported in the 2020 U.S. Census

4 Classification follows Labov (2006)

Conversational Speech Recognition (LVCSR), sponsored by the U.S. Department of Defense (Canavan et al., 1997; Kingsbury et al., 1997). Speakers were solicited by the LDC to participate in this telephone speech collection effort via the internet, advertisements, and personal contacts. The complete CallHome English corpus consists of a total of 200 unscripted telephone conversations between two native speakers of English. The LDC has pre-designated training, development, and evaluation dataset splits of 80, 20, and 100 calls, respectively. Only the 80 training calls were used for this project. The specific details of the training calls and their speakers will be described in more depth shortly.

Overall, the conversations between speakers in CallHome were completely unguided and typically between people with pre-existing friendly and casual relationships. The telephone calls were limited to a maximum of 30 minutes in length; some calls ended before the maximum was reached. Of these maximum 30 minutes, only a contiguous 10-minute period was transcribed for the training calls. The transcripts are time stamped by speaker turn for alignment with the speech signal and are provided in standard orthography (Kingsbury et al., 1997). All of the audio was recorded directly through the telephone and is contained in 8 kHz, stereo SPH format files. The SPH file is a waveform audio file created in the NIST SPHERE format, which is often used in speech recognition research.

Information was only collected for the originator of the call, the initial respondent in the participant collection process. Speaker demographic information included gender, age, years of education completed, where the speaker grew up (represented by some two-digit state code or "varied"), and the country-code or area-code of the number dialed plus the next three digits (the final four digits are encrypted as three letters).

The two-digit state code was used to hypothesize the regional dialects represented and their extent of representation in the training dataset. Because this information was not already provided in the CallHome documentation, these dialect assignments are a best guess at speaker dialect and only meant to provide a general idea of CallHome's dialect composition. Classifications were made by loosely following general regional dialect boundaries defined in the Atlas of North American English. Speakers from states which featured a dialect boundary within its borders were classified as "Mix"; likewise, for "varied" speakers. A summary can be found in Table 4.

Table 4. Hypothesized dialect region of CallHome channel 1 speakers

Dialect Region	Number of Speakers
North	40
Midland	10
South	8
West	14
Florida	1
Mix	7

Half of CallHome's primary speakers hail from northern states and likely have Northern accents. The North's representation potentially increases depending on the dialect of some "Mix" speakers from multi-dialect involved states or "varied" speakers who grew up in several states.

Unfortunately, a similar classification task cannot be done for the secondary speakers in CallHome--those receiving the phone calls--due to a lack of any explicit demographic information. All of the calls originated from North America, but 80% of the training calls were outbound to locations outside of North America. Analyzing the area and country codes reveals a diverse list of

call destinations. Most notably, 17 of the country codes are assigned to Israel; the second most popular outgoing call destination was Germany with 9 country codes matching the one assigned to the European country.

Regardless of a receiving caller's international location, all speakers in CallHome were required be native American English speakers. It is not possible to know whether the secondary speakers were originally from the United States or learned English as their first language while growing up abroad. A follow-up report on the CallHome English corpus performed manual labeling of the speakers' relationships through assessing the recorded phone call and finds that virtually all of the speakers were either family or began their relationship through work, likely in the area the primary speaker resides (Katerenchuk et al., 2018). As such, it is not unrealistic to assume that the secondary callers resemble the phonological backgrounds of their primary caller in some capacity.

Manual audit notes from the LDC do indicate that there is a presence of some non-standard speech in the CallHome audio. Specifically, 9 of secondary speakers and 4 of the primary speakers were marked as having some non-standard accent for a total of 13 non-standard speakers. Three speakers were described as either "standard-black" or "African" and two speakers were described as "Southern". The remaining of these non-standard speakers were marked as having a "Jewish" or "Yiddish" accent. Beyond these audit notes, there is no information of any speaker's race. Given the time period of CallHome's creation, processing, and release, it is likely that speakers not marked as non-standard in some way are White identifying.

In regard to previous studies, CallHome is extensively featured in the test set for modern ASR applications (Billa et al., 1999; Audhkhasi et al., 2017; Kurata et al, 2017; Saon et al., 2017). Its common inclusion as part of the Hub5-2000 Switchboard/CallHome test sets implies its

accepted utility as a metric to evaluate ASR applications. CallHome has also been used to study American English speech (Cohen Priva, 2017) and conversational processes (Horton & Gerrig, 2005).

2.5 Automatic Speech Recognition and Kaldi

Automatic speech recognition (ASR) is a technology in which a computer processes spoken speech to text, or into some semantic unit which it can utilize for further tasks. ASR systems typically involve both a language model trained on text data and an acoustic model trained on audio data. Machine learning (ML) is popular with ASR as it can provide effective modeling of the deep, dynamic structures of speech. On the other hand, ASR is an expansive, realistic problem which can aid ML development. After all, spoken human speech can quickly become a quintessential example of big, messy data.

The field of sociolinguistics has long understood that there are a myriad of dialectical varieties, influenced by race, social class, geography, and more. As discussed earlier, dialectology has shown there are measurable distinctions between speech originating from different geographical areas (Clopper et. al, 2005; Clopper and Pisoni, 2006; Labov et. al, 2006). All this being said, ASR systems perform better when the test or production data closely resembles that which it was trained with. ASR performance for English is most typically evaluated through word error rate (WER). WER is calculated as the total number of mistakes, which can be categorized as deletions, insertions, or substitutions, divided by the total number of words.

Kaldi is an open-source ASR toolkit developed in academia primarily for research (Povey et al., 2011). It has been shown to provide competitive results using state-of-the-art techniques without extensive scripting demands on the user (Gaida et. al, 2014; Morbini et. al 2013; Georgila et al., 2020). Part of the reason why Kaldi is able to provide quick and effective out-of-the-box

functionality is due to widespread recipe sharing from its active user base. A Kaldi recipe is "a set of scripts detailing steps of code execution that will enable a user to build a recognizer for some speech corpus or corpora" (Guglani and Mishra, 2018). Nonetheless, Kaldi is fully customizable, including its prepared recipes, and affords many opportunities for fine-tuning models and input. Given these benefits, it has been used in many previous ASR projects covering a wide range of topics.

Elmahdy, et. al (2013) used a Kaldi based GMM-HMM architecture to examine a transfer learning approach to ASR of an under-resourced Arabic dialect. Similarly, Menon et al. (2018) utilize Kaldi to build multilingual LSTM, BLSTM and various TDNN-based acoustic models for the under-resourced Somali language. Kaldi was used by Hermann and Magimai-Doss (2020) to build an English speech recognition system for dysarthric speakers. An even larger number of studies use Kaldi for building and improving general English ASR systems as well (for some examples, see: Graves & Jaitly, 2014; Maas et al., 2013; Snyder et al., 2018; Srivastava et al., 2014). Kaldi is even used to build other speech recognition tools. It is the base for the Montreal Forced Aligner, a tool for aligning pre-existing orthographic transcriptions to audio, typically used to create Praat TextGrids for linguistic analysis (McAuliffe et al., 2017).

2.5.1 The ASpIRE Model

ASpIRE is an nnet3 chain model trained on Fisher English augmented with impulse responses and noises. The Kaldi nnet3 is a framework for deep neural network (DNN) acoustic modeling. A chain model is a type of DNN-HMM model, which combines deep neural networks and Hidden Markov models. Specifically, ASpIRE is composed of a time-delay neural network (TDNN) and a bi-directional long short-term memory (BLSTM).

A time-delay neural network is a multilayer neural network able to perform shift invariant classification by representing temporal relationships between features through differentiating weights on input delays. The TDNN architecture is modular and incremental. It is often used in ASR because it does not require explicit segmentation of the input beforehand and precise localization of word in a speech signal is often impossible. Additional details on TDNNs can be found in Waibel et. al (1989).

An LSTM is another type of neural network which uses recurrent feedback connections and can process sequences of data. It commonly uses a memory block controlled by an input gate, a forget gate and an output gate. The blocks are recurrently connected to form the network, and when information is allowed to flow both forward and backward in time, a BLSTM is obtained. BLSTMs are well-suited to ASR because they can process long-term temporal context dependencies. More on LSTMs can be found in Sak et al. (2014).

ASpIRE was uploaded to the web and made available by Dan Povey on October 15, 2016. The README.txt file which accompanies the ASpIRE download reports a WER of 15.60 on a small hold out of training data. ASpIRE was chosen because it was trained on conversational speech and provided promising results in preliminary tests. Section 5.3 further discusses model choice and factors influencing this decision.

Again, ASpIRE was trained on the Fisher English Corpus. Fisher English was created in 2003 under the then new Fisher protocol designed to address the critical need of building robust ASR systems. Under the Fisher collection protocol, an LDC platform connects participants who typically do not know each other in a single phone call to discuss a set of predefined topics (Cieri et al., 2003). This maximizes inter-speaker variation and vocabulary breadth while also increasing formality. Not much detailed information on the demographics represented in Fisher English is

accessible, but Cieri et al. (2003) and Cieri et al. (2004) both assert that efforts were made to collect a representative sample of the United States population. Cieri et al. (2004) shows the final Fisher English corpus is largely composed of Northern and Midland speakers. Race data is entirely absent and presupposing that the Fisher English speakers are likely predominantly White identifying is not an unreasonable assumption.

CHAPTER 3

USE OF A PRE-TRAINED ASR MODEL ON AAL

3.1 Methodology

The CallHome files were downloaded in their distributed form from the LDC (Canavan et al., 1997) with assistance from the UGA Linguistics Lab. As such, all audio data was received in the original SPH file format. The transcript data was also included (Kingsbury et al., 1997).

The CORAAL files were acquired directly from the University of Oregon's Online Resources for African American Language website[5]. These files included the audio recordings and time-aligned orthographic transcription for the ATL, DCB, PRV, and ROC.

Both the interviewer and interviewee audio data are utilized for CORAAL experiments. It appears that all CORAAL interviewers are also black and originate or have close ties to the city they conducted their interviews in. As such, their inclusion should only help to provide rich additional data for each CORAAL component. All speakers on both ends of the telephone call in CallHome are used for experiments except for the three speakers marked as having "standard black" or "African" accents in the metadata.

Each corpus underwent identical preprocessing, feature extraction, and decoding procedures. These steps were applied individually such that the data from each corpus remained

5 https://oraal.uoregon.edu/coraal

separate despite the identical processing pipelines. This ensures that the distinct characteristics from each corpus remains intact and allows for clearer interpretation of the results.

3.2 Preprocessing

3.2.1 Transcript Preprocessing

Preprocessing of the ground-truth human transcriptions was tailored to the ASpIRE model's expectations in order to minimize the WER. Discrepancies in a model's lexicon and the transcripts' conventions will always inflate WER regardless of a model's ability to correctly recognize a word. Additionally, cleaning the transcripts ensures more consistency across each dataset examined.

To begin with, flags for redacted words (e.g., identifying names, addresses), nonlinguistic markers (e.g., coughs, laughter), and unintelligible audio content were removed. Dashes or hyphens were substituted with a singular space. Additionally, all punctuation besides apostrophes was scrubbed and characters were converted to lowercase.

After examining the model output, a decision was made to instead replace notes of laughter in the transcript with "[laughter]", the form which the ASpIRE model expects. While ASR systems typically ignore nonlinguistic markers like laughter, ASpIRE, again, was augmented with impulse responses and is therefore capable of detecting and reporting laughter in its hypothesized transcript. Additionally, some exceptionally common errors caused by a mismatch in transcript conventions were converted to the ASpIRE expectation: standalone instances of *dc* (as in "Washington D.C." or "the D.C. area") were replaced with *d._c.*, *cause* with *'cause*, and *mm hm* with *mhm*. Overall, the new four changes decreased WER across each corpus from as little as 0.07 to as much as 2.87. The original, uncleaned transcripts were never given to the ASpIRE model for

scoring, but the associated WER of such an iteration would undoubtedly be higher by a substantial amount.

Of course, there are other conventions found in the CORAAL and CallHome transcripts which could be amended to better match those used for the Fisher English transcripts. For example, not all acronyms were converted to the form expected by the ASpIRE model and there were a few instances of purposely misspelled words in the CORAAL transcript. Misspellings were done to "accurately account" for the actual pronunciations produced by the speaker (Kendall and Farrington, 2020). A common misspelling was the transcription *aks* for *ask*; another example of a purposefully misspelled word in the transcription was *thame* for *same*. However, the effect of these problematic words on the overall WER would be negligible given their rarity and the size of each dataset. As such, no further effort was taken to modify the ground-truth human transcripts.

3.2.2 Audio Preprocessing

At this stage, no alterations were necessary for the CORAAL audio. However, the same cannot be said of the CallHome audio. Kaldi requires single-channel files. As such, CallHome's SPH files were converted to two different .wav format files, one for each channel on the original file, using the LDC's sph2pipe tool (Graff et al., n.d). Only the ten-minute portion transcribed in the LDC transcripts is included in the resulting .wav files in order to reduce processing time and storage requirements.

3.2.3 Preparation for Kaldi

Certain files need to be created for Kaldi specifically. For each dataset, a "text", "segments", "utt2spk", and "wav.scp" file was created by manipulating the non-utterance data associated with the now-cleansed transcripts. Table 5, with the structure of each file and an example line, is found below. Principally, the Kaldi files require certain information arranged in

file-specific ways. The required information included speaker IDs, utterance IDs, file IDs, file paths, utterance start and end times in their respective audio file, and the text transcription of the utterance. An explanation of how each ID was formed follows.

Table 5. Structure and example lines for files required to run Kaldi

File	Structure	Example
text	<utterance-id> <text_transcription>	ATL_se0_ag1_f_01-n0105 seventh grade you start having all the ratchetness
segments	<utterance-id> <file-id> <utt-start> <utt-end>	ATL_se0_ag1_f_01-n0105 ATL_se0_ag1_f_01_1 454.4563 457.2352
utt2spk	<utterance-id> <speaker-id>	ATL_se0_ag1_f_01-n0105 ATL_se0_ag1_f_01
wav.scp	<file-id> <full_path_to_audio_file>	ATL_se0_ag1_f_01_1 /home/norad/coraal/ATL/ATL_audio_2020.05/ATL_se0_ag1_f_01_1.wav

Each speaker is identified by a unique speaker ID. For the CORAAL datasets, the speaker ID follows the text found in the "Spkr" column of the transcripts. CallHome speaker IDs were formed by affixing the text found in the "turn" column of its transcripts with the numbers from the file name, separated by a singular underscore to match the CORAAL convention. In both PRV and CallHome, interviews occasionally featured a third or even fourth speaker, albeit always in very limited capacities. PRV distinguishes all of these speakers by labeling them as "Misc". CallHome adds a "1" (or sometimes "2") to the letter used for that channel (e.g "B1"). The same process was applied to these speakers.

Furthermore, every utterance is assigned a unique utterance ID. Partwise, the utterance ID is a concatenation of the speaker ID and a newly generated alphanumeric code that contains at least one letter and four digits. Digits increased sequentially throughout a speaker's utterances such

that their first utterance in the transcript would be “fk0001” and their last “fk3467”. Within datasets, these alphanumeric codes are unique.

The fileID is simply the name of the file which has remained unchanged since download. The file path is the absolute path to the .wav audio file. The utterance start and end times are directly extracted from the transcript. In the case of CallHome, times are shifted to match the newly isolated ten minute .wav files.

3.3 Kaldi and ASpIRE

3.3.1 Configuration

Kaldi version 170a1fc was used for all experiments (Povey et. al, 2011). The ASpIRE model without the precompiled HCLG (decoding graph) was obtained directly from the Kaldi website. A precompiled HCLG is available from the same source and compiling the HCLG on the local system produces the same graph. All experiments were performed on a virtual machine allocated 8 processors, 12gb RAM, and running the Ubuntu 20.04 operating system.

3.3.2 Feature Extraction

Following various examples from existing Kaldi recipes and online resources, a shell script was written to carry out the various necessary steps to apply the ASpIRE model to the new data.

In this, 40 Mel-frequency cepstral coefficients (MFCCs) were extracted for each utterance. MFCCs are a very commonly used feature in audio research due to their ability to assess pathological speech and represent frequency regions audible to the human ear (Khan, 2014). Due to the Fisher corpus’ 8 kHz nature, the CORAAL data had to be downsampled to 8 kHz. Kaldi allows this step to occur in conjunction with the MFCC calculations through its extraction pipeline. CallHome’s sampling rate is already 8 kHz, so no down- or upsampling is required.

Occasionally, some utterances would be discarded due to insufficient length to extract meaningful MFCCs. A summary of the resulting number of speakers, utterances, and words after feature extraction can be found in Table 6.

Table 6. Number of speakers, utterances, and words in each dataset after preprocessing and feature extraction

Dataset	Number of Speakers	Number of Utterances	Number of Words
ATL	14	15211	91806
DCB	52	79213	508430
PRV	21	30438	153229
ROC	16	22323	136900
CallHome	164	19519	161817

3.3.3 Decoding and Scoring

Online decoding using the Kaldi nnet3 decode script was performed for all of the datasets. Kaldi's online decode script automatically computes derived variables for iVector extraction. An iVector is "a vector of dimension several hundred...which represents the speaker properties" (Povey, n.d). Povey asserts that iVectors provide the model with as much as it needs to know about speakers and will increase accuracy. In Kaldi, its estimation is Maximum Likelihood, involving Gaussian Mixture Models.

The acoustic scale is set to 1.0, as the default (0.1) is not suitable for a chain model like ASpIRE. Additionally, the post decode acoustic scale is set to 10.0 which scales the acoustic probabilities by 10 after decoding so the regular scoring script will function properly in the chain system. These are the recommended values for Kaldi chain models.

The model decode results were then scored against the ground-truth human transcripts using Kaldi's built-in scoring script. The scoring process iterates through the decode results with different language model weights (ranging from 7 to 20) and word insertion penalties (0.0, 0.5, or 1.0). The language model weight is the inverse of the acoustic scale and is the amount by which the language model probabilities are scaled. It affects the influence the language model exerts on the produced transcript in conjunction with the acoustic model. The word insertion penalty is a fixed value added to each token in the decode results, penalizing insertion errors.

Total run time in Kaldi was approximately 31 hours for all five datasets.

CHAPTER 4

RESULTS

Table 7 displays the WER for each dataset, listed from lowest to highest, along with a summary of the types of errors made. CallHome was the best performing dataset with a WER of 23.99. Following, ROC had a WER of 27.96, DCB 36.99, ATL 42.36, and PRV 50.53. For every dataset, the best performing WER was produced with a language model weight of 8.0 and a word insertion penalty of 0.0.

About half of the errors for each dataset were substitution errors, suggesting that the model detected words but was unable to correctly identify them as opposed to extraneous artifacts in the audio.

Kaldi provides a more detailed scoring summary for each dataset which includes individual words' error types and quantity, errors by speaker, and by utterance. These results will be used in further analysis in the next section.

Table 7. Overall WERs and error summary for each dataset

Dataset	Word Error Rate	Error Summary
CallHome	23.99	38820 / 161817 insertions: 3877 deletions: 12728 substitutions: 22215
ROC	27.96	38278 / 136900 insertions: 3340 deletions: 13521 substitutions: 21417
DCB	36.99	188079 / 508430 insertions: 13655 deletions: 67063 substitutions: 107361
ATL	42.36	38887 / 91806 insertions: 3029 deletions: 13437 substitutions: 22421
PRV	50.53	77429 / 153229 insertions: 4534 deletions: 29476 substitutions: 43419

CHAPTER 5

EFFECT OF AAL AND REGIONAL VARIATION ON ASR PERFORMANCE

Kaldi and the ASpIRE model performed worse on all of the CORAAL datasets compared to CallHome. Within the CORAAL corpus, the WER for PRV is nearly twice that of ROC. These results betray underlying biases in the ASpIRE model. In order to explore these potential biases, select AAL and regional phonological phenomena are examined.

5.1 Phonological Sources of Recognition Errors

A sample of phonological characteristics of AAL are chosen from Table 1 to be examined in context of the ASpIRE model performance:

- reduction of word-final consonant clusters that end in [t] or [d],
- [θ] as /t/ or /f/ and [ð] as /d/ or /v/,
- vocalization of post-vocalic [l],
- and deletion of word-final, postvocalic [r].

Certain restrictions on phonological context were placed in order to most likely capture the AAL realization of these phonemes and phenomena. Table 8 shows a sample list of words which contain the phonemes in the relevant phonological context.

Table 8. Example words containing AAL associated phonemes.
The relevant phonological context is highlighted in the word's CMUdict pronunciation mapping.

Word	CMUdict Pronunciation
Word-final [t, d] consonant cluster	
best	B EH S T
called	K AO L D

left	L EH F T
Voiceless dental fricative [θ]	
both	B OW TH
thank	TH AE NG K
thing	TH IH NG
Voiced dental fricative [ð]	
that	DH AE T
they	DH EY
these	DH IY Z
Postvocalic [l]	
cool	K UW L
help	HH EH L P
told	T OW L D
Word-final postvocalic [r]	
four	F AO R
more	M AO R
sure	SH UH R

Notably, all of the features selected to represent AAL concern only consonants. Section 2.2 showed that regional dialects are more characteristically defined by their vowel spaces. As such, to investigate the effect of regionality on ASR performance, six different vowels are explored:

- the diphthong AY (/aɪ/, *ride*),
- the diphthong EY (/eɪ/, *bait*),
- the front-high vowel IY (/i/, *bee*),
- the mid-central vowel AH (/ʌ/, *bus*),
- the low-back rounded vowel AO (/ɔ/, *thought*),
- and the low-back unrounded vowel AA (/ɑ/, *bot*).

All of these vowels are affected in some way by the NCS, SVS, or AAVS. Again, in the SVS, AY is monophthongized, EY moves towards the space traditionally occupied by EH, and IY

is backed and lowered. In the NCS, AH is backed, AO is lowered, and AA is fronted. A sample list of words containing these vowels is seen in Table 9. These vowels were chosen because there is not substantial overlap expected between the three potential shifts represented in the data. In other words, Northern, Southern, and African American speakers are predicted to realize these vowels in different ways that do not necessarily conflict. Nonetheless, AA is also affected by the AAVS in the same way as the NCS. This dual impact may be interesting to further investigate in the data.

Table 9. Example words containing regionally variable vowels.
Relevant vowel is highlighted in the word's CMUdict pronunciation mapping.

Word	CMUdict Pronunciation
AY (/aɪ/, *ride*)	
by	B AY
child	CH AY L D
drive	D R AY V
EY (/eɪ/, *bait*)	
days	D EY Z
eight	EY T
grade	G R EY D
IY (/i/, *bee*)	
eat	IY T
free	F R IY
need	N IY D
AH (/ʌ/, *bus*)	
bus	B AH S
comes	K AH M Z
front	F R AH N T
AO (/ɔ/, *thought*)	
born	B AO R N
lost	L AO S T
wrong	R AO NG
AA (/ɑ/, bot)	
far	F AA R

got	G AA T
job	JH AA B

Of course, there are other features of both AAL and regional dialects which are not examined here yet that may have an influence on the model's performance. Common and distinguishable features of each dialect were chosen to provide a general overview of how these non-standard features may affect ASR.

In order to examine a somewhat representative sample of phonological realizations from each region, only words which appear in each dataset more than ten times are included in the analysis. Ten is the minimum number of times a word must appear in the Atlanta subset, the smallest dataset of the five, to constitute at least 0.0001% of the transcript without rounding. There is a total of 479 words which meet this criterion. The WER for these words is reported in the second column of Table 10.

Monosyllabic words may provide a clearer depiction of only the phenomena of interest. When performance for only single syllable words is considered, we see a rise in WER across all datasets as compared to the 479 common words. Likely, the additional phonetic information provided by longer, multisyllabic words assists the model in selecting a correct hypothesis. In order to best isolate the phoneme of interest, only monosyllabic words are included in this analysis. 333 words are both monosyllabic and occur in each dataset more than ten times. The full list of the 333 common monosyllabic words can be found in Appendix A. The WER for these words is reported in the third column of Table 10.

A pairwise t-test across all the corpora with Bonferroni adjustment on the *p*-value yields that the average WERs for common monosyllabic words are significantly different ($p \leq 0.05$) from one another for all datasets with the exception of DCB and ATL ($p = 1.0$) for these common

monosyllabic words. The high *p*-value does not imply that the speakers of DCB and ATL are identical in their phonology, but rather that the model performs equally on the speech samples. Further pairwise t-tests are performed with the AAL and regional subsets and will be discussed in the following sections.

In order to determine if a word contains a phoneme of interest, the Carnegie Mellon University Pronouncing Dictionary[6] (CMUdict) is used to systematically select words. The CMUdict is an open-source pronunciation dictionary for North American English that contains over 134,000 words and their mapping(s) to pronunciations in the ARPAbet phoneme set. A list of the unique words found across all of the CallHome and CORAAL data is given to its Lexicon Tool[7] in order to produce a filtered pronunciation dictionary. The pronunciations from this filtered dictionary guide word selection per phoneme. For example, to compile the list of words which potentially undergo deletion or vocalization of postvocalic [l], the pronunciations from the CMUdict are referenced to find words which contain an L after a vowel. Similarly, words which contain the vowels AA or AO in their CMUdict pronunciation mapping are selected for those categories. In cases where a word has multiple pronunciations and one or more of those pronunciations contains a phoneme of interest, the word is excluded entirely if the conflict would result in unclear or variable realizations. For example, *the* has both the mappings DH IY and DH AH. *The* is included in the list of words for [ð] because DH is common to both CMUdict pronunciations, but *the* is excluded from both the IY and AH word lists.

6 http://www.speech.cs.cmu.edu/cgi-bin/cmudict
7 http://www.speech.cs.cmu.edu/tools/lextool.html

Even still, many words exhibit several phonological features of interest in a single pronunciation mapping, such as the word *felt* which contains both a postvocalic [l] and a word-final consonant cluster that ends in [t]. Because we are examining these dialectal phenomena in a word recognition context, we cannot simply isolate the individual phonemes for study. While it is easy enough to exclude words which contain more than one consonant related feature, it would be unproductive to then limit those words to only having vowels which have resisted any observed change due to the heavy overlap in vowels affected by the SVS, NCS, or AAVS. In practice, the inclusion of words with overlapping features of interest does not do much to skew the data one way or another.

Table 10. WERs for (1) all words which appear in each dataset more than ten times and (2) monosyllabic words which appear in each dataset more than ten times

WERs for Common Words		
Dataset	n≥10	n≥10 and monosyllabic
CallHome	16.86	18.84
ROC	21.42	24.35
DCB	31.63	34.51
ATL	32.80	35.67
PRV	46.31	48.13

For the remainder of this thesis, the words discussed and used for calculations are those which are monosyllabic, occur in every dataset, and occur more than 10 times, unless stated otherwise. Table 1 presents the WERs of all words, these common monosyllabic words, and the

common monosyllabic words which contain either the AAL associated phonemes or the regionally variable vowels.

5.1.1 AAL Association

There is a total of 105 words which encase one (or more) of the phonological contexts where an AAL feature may be present. These will be known collectively henceforth as the "AAL words" for brevity. The average rate at which the ASpIRE model correctly identifies these words is aggregated in Table 11 by phoneme and subcorpus. The voiceless and voiced dental fricatives are separated into two categories because their AAL realizations take on different forms and their differences in WER are quite large.

Table 11. Average rate of correct identification of AAL words by subcorpus

Average % Correct	CallHome	ROC	DCB	ATL	PRV
Word-final [t, d] consonant cluster	0.8265	0.7663	0.6617	0.6395	0.5039
[θ]	0.8810	0.7672	0.7158	0.6328	0.5596
[ð]	0.7811	0.7207	0.6108	0.5632	0.4624
Postvocalic [l]	0.8055	0.7129	0.6122	0.5757	0.4561
Word-final postvocalic [r]	0.7797	0.7470	0.6257	0.6567	0.4378
Overall Average	0.8148	0.7428	0.6452	0.6136	0.4840
WER Average	0.1852	0.2572	0.3548	0.3864	0.5160

The average WER difference for the voiceless dental fricative [θ] is the smallest between ROC and the other CORAAL components. The ASpIRE model performs, on average, 21.22%

better on these words than all of CORAAL. The boxplots of the rate of correct identification of these words are showcased in Figure 1. Average % correct for words containing [θ], reveals that all of the CORAAL datasets have similar ranges, although ROC's lowest point is an outlier in the word "thank" at 36.37%. In contrast to the wide spread of the CORAAL data, CallHome's correct rates have much less deviation. It's likely that the realization of [θ] in CallHome is more consistent than in CORAAL.

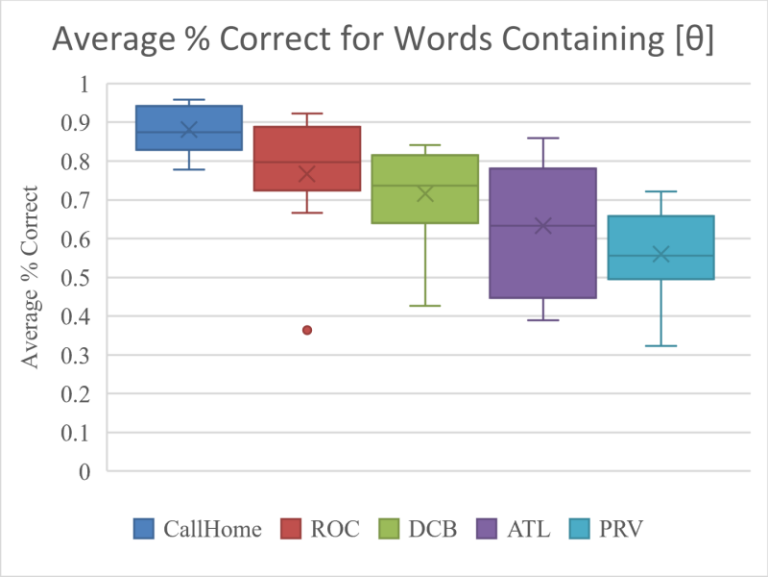

Figure 1. Average % correct for words containing [θ]

The same is true, although to a lesser extent, for words with word-final consonant clusters ending in [t] or [d], the plots of which can be found in Figure 2. Farrington (2018) found the duration of vowels before underlying [d] in consonant neutralized contexts is significantly longer than for [t]. This change in vowel duration may explain the larger variance in the CORAAL data, as the model struggles with not only the potential consonant cluster reduction but also an unexpected vowel duration.

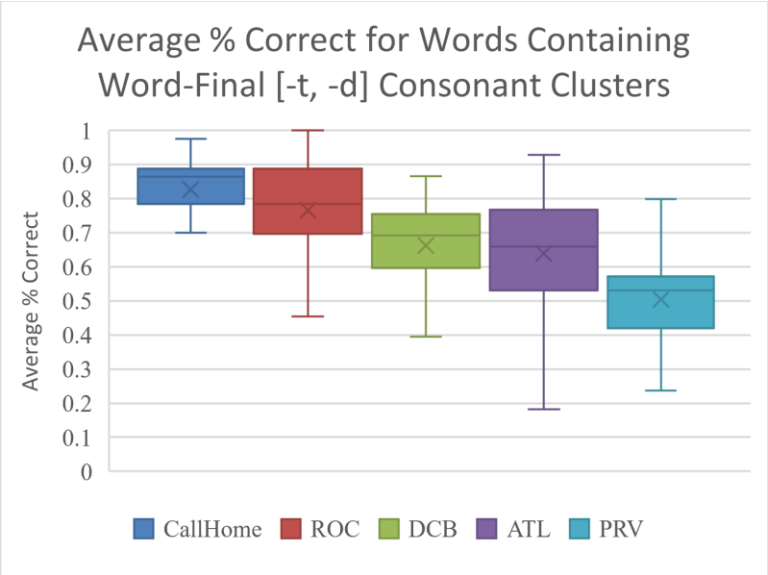

Figure 2. Average % correct for words containing word-final consonant clusters ending in [t] or [d]

Overall, CallHome's WER is 5.51 lower than ROC for common monosyllabic words, as can be calculated from Table 10. For the AAL phonological features, the gap in WER widens in all categories except word-final postvocalic [r]. CallHome and ROC only differ by 3.26 in their WERs for word-final postvocalic [r]. Rochester is a securely Inland Northern city and as such, its speakers are likely much more rhotic than the other CORAAL speakers. Figure 3 displays the boxplots for these word-final postvocalic [r] words across all of the datasets. The upper quartiles for CallHome and ROC are very similar, and ROC's lower quartiles lag behind only slightly. On the other hand, there is a much more noticeable gap between ROC and DCB, ATL, or, most extremely, PRV.

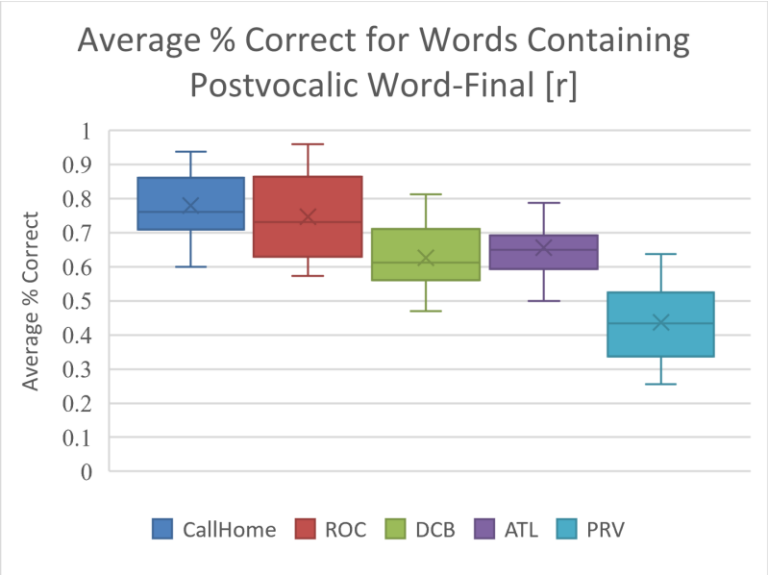

Figure 3. Average % correct for words containing postvocalic word-final [r]

While non-rhoticity was once widespread across the United States, rhotic accents have increasingly become more rampant and thus more associated with SAE (Millar, 2012). This is even true for the Southern United States, but primarily only for white speakers (Thomas, 2008). These linguistic trends explain why words containing word-final postvocalic [r] are more often recognized for ROC than, say, PRV, a deeply Southern and Black community. Indeed, Hinton and Pollock (2000) found that there was a notable difference in productions of vocalic and postvocalic /r/ across African American speakers from Davenport, IA and Memphis, TN. Davenport is another Inland Northern city while Memphis is a Southern city, mirroring the geographic and dialectal associations of Rochester and Princeville. As such, rhoticity is an example where some African American speakers may deviate from the prescribed AAL behavior in favor of regional or local production patterns.

It is clear that the ASpIRE model performs worse on the CORAAL datasets compared to the CallHome dataset. When only examining the AAL words, the WER for each of the CORAAL

datasets increases from that of the common monosyllabic words. In contrast, CallHome's WER actually decreases by 0.32 in the same words. These differences can be found by comparing the WERs for each dataset in Table 8 and Table 7 and are more clearly highlighted in Table 1. Words which contain AAL associated phonemes are recognized less well relative to all of the common monosyllabic words. This effect is not happenstance and is rooted in the non-standard realizations of these words by Black speakers using AAL. However, AAL usage and race alone are not sufficient to explain the drastic differences in WER across the CORAAL datasets. Following the revelation that ROC speakers are more rhotic than their non-northern CORAAL counterparts, the geography of the speakers will be further investigated in the next section.

5.1.2 Regional Association

Of the 333 common monosyllabic words, 152 contain one of the six regionally variable vowels listed earlier. The average rate at which the ASpIRE model identifies these words is aggregated in Table 12 by phoneme and subcorpus. These will be known collectively henceforth as the "regionally variable words" for brevity.

Table 12. Average rate of correct identification of regionally variable words by subcorpus

Avg % Correct	CallHome	ROC	DCB	ATL	PRV
AY (/aɪ/, *ride*)	0.8254	0.8056	0.6618	0.6699	0.5249
EY (/eɪ/, *bait*)	0.8154	0.7819	0.6769	0.6772	0.5382
IY (/i/, *bee*)	0.8578	0.8234	0.7093	0.6275	0.6083
AH (/ʌ/, *bus*)	0.8298	0.7693	0.6639	0.6663	0.5413
AO (/ɔ/, *thought*)	0.8245	0.7750	0.6694	0.6500	0.5089
AA (/ɑ/, *bot*)	0.7875	0.7847	0.6522	0.6880	0.5754
Overall Average	0.8234	0.7900	0.6722	0.6632	0.5495
Overall WER	0.1766	0.2100	0.3278	0.3368	0.4505

Unlike with the AAL words, the overall WERs for each dataset decrease in Table 12 from Table 10. Again, this difference can be clearly seen in Table 1 as well. In other words, the words containing the regionally variable vowels are better recognized than words containing AAL associated phonemes. This is not a surprise as the words selected are no longer exclusively those where we expect to find AAL-based deviation from the standard. The ASpIRE model was trained with some variety in vowel realizations, but AAL realizations of consonantal patterns in speech are likely absent from the acoustic model. Nonetheless, there is still a strong upward trend in the WER as the speakers become increasingly Southern.

The pairwise t-tests with Bonferroni *p*-value adjustment for the WERs on the AAL words revealed each dataset was significantly different besides DCB and ATL ($p = 1.0$). However, the pairwise t-tests for WERs on the regionally variable words produced *p*-values of 1.0 for DCB and ATL and 0.22 for CallHome and ROC. Importantly, the WERs for the CallHome and ROC regionally variable words are not significantly different. The WERs found in Table 12 for CallHome and ROC are not exactly the same, but they are more similar compared to those found in Table 11 for CallHome and ROC. While the non-significant *p*-value does not entail that the vowel realizations of the speakers in CallHome and ROC are the same for these words, it indicates the model is performing similarly in general. This similar performance is further scrutinized next.

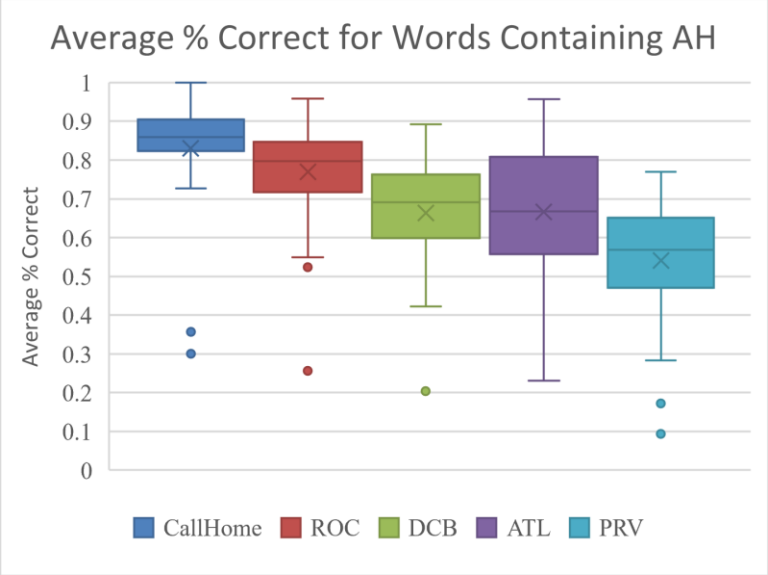

Figure 4. Average % correct for words containing AH

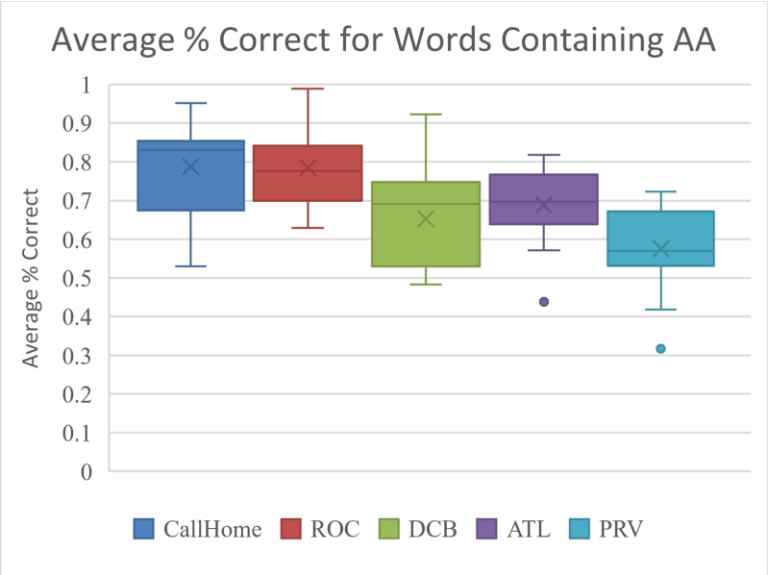

Figure 5. Average % correct for words containing AA

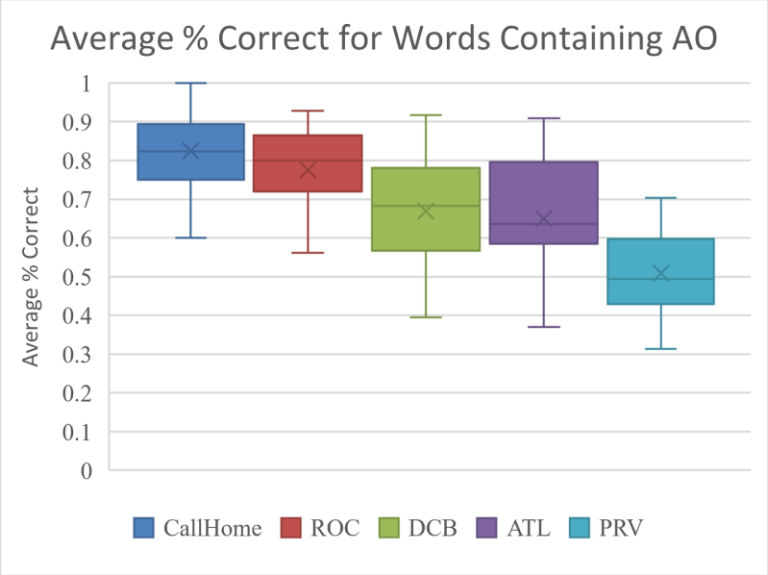

Figure 6. Average % correct for words containing AO

The two largest gaps in WER between CallHome and ROC are for AH and AO, vowels that are backed and lowered in the NCS, respectively, but are not implicated by the SVS or AAVS. In contrast, AA is fronted in both the NCS and AAVS, and the difference in WER for AA between CallHome and ROC is the smallest of all of the vowels selected. As such, it's possible that African

Americans in Rochester do not engage in the NCS in their speech, but the AAVS leads to AA fronting similar to that found in CallHome speakers. One explanation for why ROC is much closer in WER to CallHome for AA words is that Rochester speakers undergo a dual influence from both the NCS and AAVS which more intensely produces AA fronting. The plots for all three of these vowels are in Figure 4, Figure 5, and Figure 6.

In terms of the vowels involved in the SVS but not in the NCS, the WERs for ROC and CallHome are more similar compared to the WERs for the NCS-affected AO and AH vowels. As seen in Figure 7, Figure 8, and Figure 9, the boxplots for ROC and CallHome largely overlap. It is expected that neither the speakers in CallHome nor ROC would modify these vowels as they would not have any Southern pressure to do so. On the other hand, the WERs of the remaining CORAAL datasets are staggered despite these vowels' lack of association with the AAVS. The model's performance on these words indicates that Black speakers in the South do participate in the SVS, although to varying degrees depending on the city they reside.

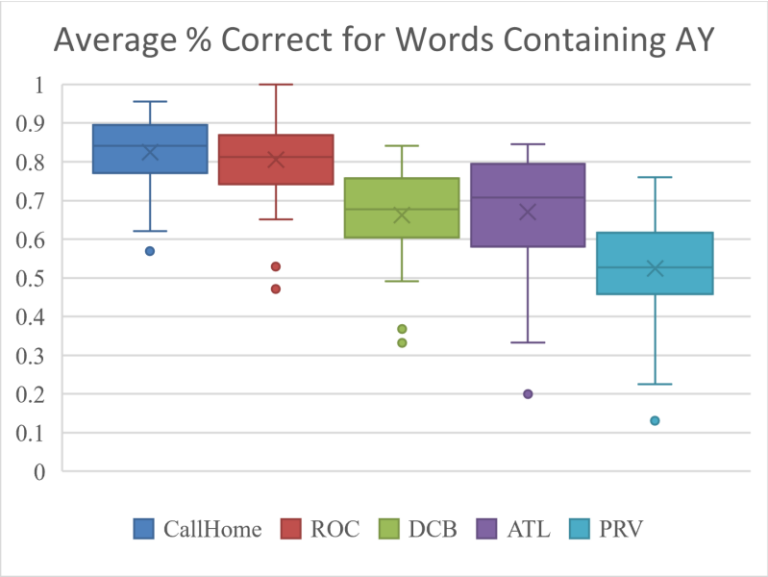

Figure 7. Average % correct for words containing AY

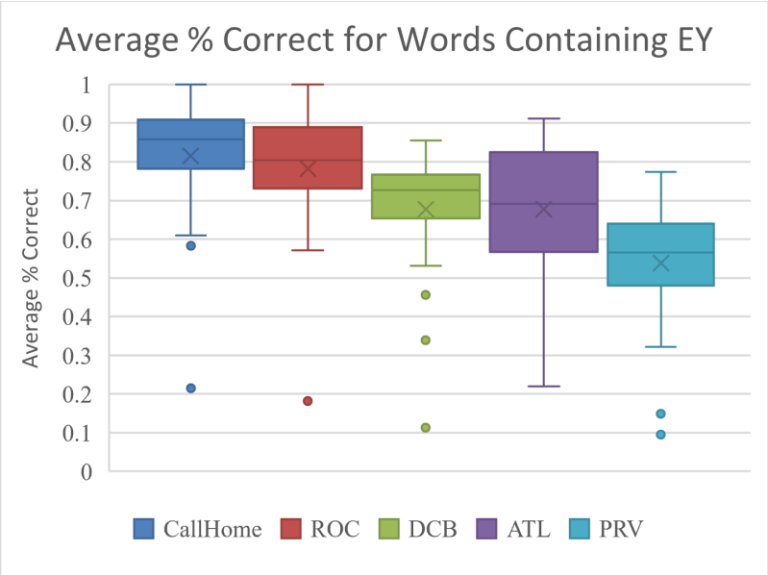

Figure 8. Average % correct for words containing EY

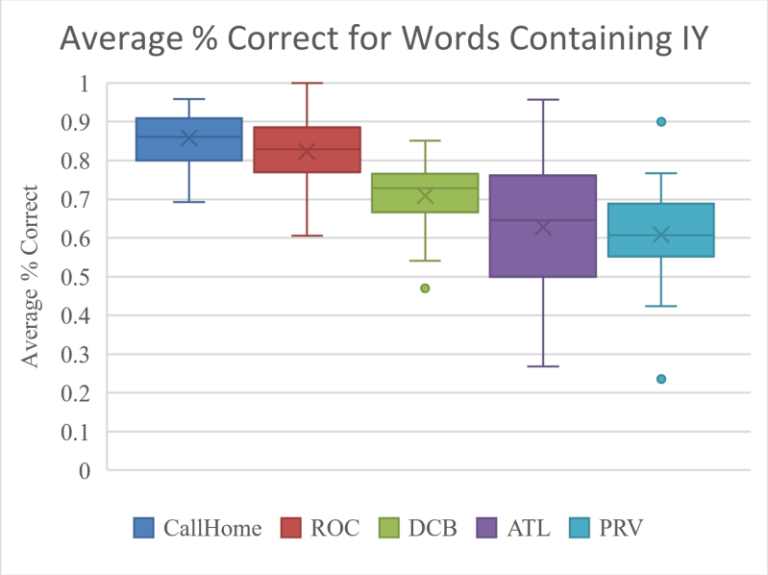

Figure 9. Average % correct for words containing IY

The diphthong AY is a good example of this trend. ASpIRE suffers a much higher WER for PRV, the dataset expected to have the most glide deletion of AY as Princeville is a distinctly Southern town. In fact, PRV's maximum individual score just barely breaches DCB's upper quartile, as seen in Figure 7. Atlanta is also a Southern city, but its AY WERs paint a much more

complicated image of its speakers' glide weakening. Labov et al. (2006) strongly claims Atlanta speakers are an exception to the South's glide deletion (amongst being resistant to other SVS phenomena). However, Kretzschmar (2015) finds that Atlanta speakers do show AY glide weakening. Furthermore, Kretzschmar (2015) finds that while there is evidence of the EY and EH reversal typically associated with the SVS in Atlanta speakers, this does not carry onto IY and IH. Labov et al. (2006) does outline that the SVS can occur in three stages: AY glide deletion, then reversal of EY and EH, and finally reversal of IY and IH. It's possible that the extreme range captured by the ATL AY, EY, and IY correct word rates reflects its speakers' transitional and varied relationship with the SVS.

Additionally, Adams (2009) found that vowel length is consistently shorter in AAL than in the NCS for tense vowels in the IY/IH and EY/EH tense/lax vowel pairs. Thus, while both AAVS and the NCS include AE raising, only the NCS experiences a subsequent vowel chain shift because of the length differences found in AAL (Adams, 2009). Adams' findings do more to explain the overall larger variances found in the CORAAL IY correct word rates. Shorter vowel lengths may provide the model with less phonetic information with which to correctly identify words. It's also possible that any divergence from the ASpIRE acoustic model's expectations of the IY realization will negatively affect the correct identification rate. This difference in vowel quality coupled with Southern speakers' backing and lowering of EY and IY results in higher WERs for PRV.

Washington D.C. does not fit squarely into any one regional dialectal area because its speakers have been shown to adopt patterns from the Northern, Midland, and Southern regions (Labov et al., 2006). Like Atlanta, this is due to its continuous and strong population movement into the city. ATL speakers hail from a much more diverse set of locations, spanning from New

York to Alabama. In fact, all but three of the interviewed Atlanta speakers in the CORAAL component are transplants. Similarly, only 11 of the speakers in DCB have never lived anywhere else. In contrast, almost every speaker in ROC and PRV had always lived in Rochester or Princeville, respectively, or came from a nearby state/area with the same dialect categorization. Nonetheless, the DCB speakers, if not D.C. native, are primarily from Maryland or Virginia. Thus, the fact that DCB's correct word rates for these vowels shows less variation than the same subset of ATL data may be due to the DCB speakers' denser concentration of geographical origin compared to ATL. Consequently, the ATL and DCB corpora feature a wider range of phonological backgrounds as compared to the other CORAAL components. Additionally, some speakers in the ATL and DCB components did originate from other Southern states.

Now, we would expect words containing vowels involved in only the NCS from non-Northern speakers be better recognized by the model because their pronunciations would presumably be closer to the prescribed position of the vowels. Of course, the ASpIRE model does not perform better on DCB, ATL, nor PRV over CallHome or ROC for AO or AH. This is due to the fact that the ASpIRE model was not trained on textbook pronunciations of words, but conversational speech from primarily Northern and Midland speakers. In fact, this is likely why PRV is the most poorly recognized across the board. Black, Southern speakers are not well represented in Fisher English. Then DCB and ATL see less than optimal performance compared to CallHome and ROC due to their mixed variety of dialects and transitional realizations of vowels which would similarly not be adequately reflected in Fisher's Northern speakers.

5.2 Lexical Sources of Recognition Errors

As seen in Table 10, the WER for each dataset decreases from its overall WER when words featured in each dataset are used for the calculations. This likely stems from the fact that words

that appear in all of the five datasets are more common words in general and thus more likely to be recognized by the model in general.

However, there are some words which are misrecognized at much higher rates than the average word, many of them coming from one of the CORAAL datasets.

In total, there were 3642 unique words found in the transcripts for CallHome and CORAAL that were not in the ASpIRE lexicon[8]. As such, not a single instance of any of these words was correctly identified by the model. While the number of times these words appear in the transcript is minimal relative to the overall size of each dataset, they expose a deficiency in the ASpIRE lexicon. Notably, some words have roots in or ties to the African American community, such as "swag", "phat", or "freaknik". These are generally considered slang terms, but their use has been widespread in the African American community for decades and have crossed into the mainstream American vocabulary through hip-hop music or fashion brands like Baby Phat.

Additionally, the words "Princeville", "Obama", and "Eminem" were not found in ASpIRE lexicon, but other proper nouns like "Atlanta", "Romney", or "Tupac" were. The topics used to guide the original Fisher English conversations were purported to maximize vocabulary coverage, but these conversations inevitably become a relic of their time. Obama may not have been a household name during the time of Fisher English's composition, but his exclusion from the ASpIRE lexicon marks the potential need for continuous upkeep of an ASR model's vocabulary. Furthermore, the data captured in even semi-guided conversations may not be adequate to capture not only a large vocabulary but a representative vocabulary. The topics used to frame conversation

[8] Some of these instances were not fully formed words, but partial words which were not removed in the preprocessing steps.

during the creation of Fisher English evidently did not elicit words common to the Black community.

Another stand-out example of the ASpIRE lexicon's failure to meet the reality of AAL is the n-word. For whatever reason, the n-word with the "-a" ending is not found in the ASpIRE lexicon, but its rhotic counterpart (the n-word with the "-er" ending) is. In the CORAAL transcripts, the rhotic n-word (both singular and plural) is only found a total of five times, all in DCB; it makes one appearance in CallHome. Over these six instances, ASpIRE correctly hypothesizes the rhotic n-word 100% of the time. On the other hand, the non-rhotic n-word (in all forms) appears 454 times across ROC, DCB, and ATL and is never correctly identified. Table 13 displays the utterances and some context for the few instances of the rhotic n-word found in the CallHome and DCB data.

Table 13. Utterances containing the rhotic n-word. Surrounding utterances are included to give additional context to the use of the rhotic n-word.

Dataset	File ID	Utterance
CallHome	4077_B	and um francis was very worried about me because she came home two weeks before you know f um because she just was acting like a total nigger child uh oh hope the research didn't hear that [laughter]
DCB	DCB_se1_ag2_m_01	it shouldn't be used nigger nigga whatever it shouldn't be used a degrading word
DCB	DCB_se1_ag4_m_01	a whole class of niggers ... he just called us a whole class of niggers you know and t hat's that was another experience with racism so-
DCB	DCB_se2_ag2_f_02	We're walking, literally, across the street.

		Speeds past us within- closer than me and you are right now, almost hit us, and yells out the window, niggers. And I had never been in fear of my life until that very point in time. Cause I'm like, if they were to pull over, there's nothing we can do right now to defend ourselves.
DCB	DCB_se2_ag2_f_02	Apparently, these Caucasian relatives decide that they're not gonna accept that my great grandmother is getting married to a black man, and she has biracial children. So they tell her to get her nigger man and her porch monkeys off their property. And that was the extent of the relations with our Caucasian family.

While a simple “fix” would be to alter every instance of the n-word to have the orthographically rhotic ending during preprocessing, this would fundamentally alter the intention and representation of the word by the speakers. While the n-word unequivocally finds its roots as a racial slur, the rhotic n-word has historically been used by those in positions of privilege, typically white, as an act of hate against Black people. The examples of the rhotic realization in Table 13 are evidence of such. In contrast, the n-word with the -a ending is a more casual form of the word, typically used by African Americans to refer to each other informally (Yancy, 1995). As such, the weight these two different words hold is immense and their representation in the ASR output can psychologically affect the people using these ASR systems (Sue, 2013; Williams et al., 2016).

It’s possible that the transcription conventions for Fisher English defaulted to record every instance of the n-word with the rhotic pronunciation. Alternatively, the speakers in Fisher English

truly may have only used that variant. It would be interesting to analyze the original Fisher English transcripts used to train ASpIRE in order to explore the contexts in which the n-word appears and why the rhotic representation of the word is the only one found in the lexicon. In either case, there are racial biases at hand which negatively affect an ASR system's performance on Black speech when trained on Fisher English, including within its lexicon.

Regionality is also relevant to the lexical errors made by the ASpIRE model. The word "y'all" is a distinctively Southern term used to reference groups of people or address others. It does in fact exist in the ASpIRE lexicon, but of its 636 instances across all of the datasets only 27 were correctly identified. Likewise, the word "ain't", an informal contraction of "am not" or "is not", is also found in the ASpIRE lexicon, but is only correctly recognized 121 times out of 871 occurrences across all of the datasets. Unsurprisingly, the vast majority of the occurrences of "y'all" and "ain't" are not from CallHome or ROC. A detailed breakdown of the total number of occurrences and the number of times the model correctly identifies these words is found in Table 14.

Table 14. Total instances and total instances correctly recognized for words "y'all" and "ain't" by subcorpus

	y'all		ain't	
Dataset	Total Instances	Total Correct	Total Instances	Total Correct
CallHome	2	0	14	3
ROC	25	2	22	4
DCB	476	19	479	54
ATL	88	4	209	46
PRV	45	2	147	14

Sum	636	27	871	121

Despite these words being found in the ASpIRE lexicon, their low rates of correct identification suggest that they are assigned low weights in the language model. Likely, one or two Fisher English speakers used the words, but not in a large enough quantity for the model to deem these words significant. In contrast, we find these words to be much more common in Southern vernacular. As such, Southern speakers suffer poorer ASR performance for their use of these words.

5.3 Other Factors Affecting Model Performance

Some preliminary experiments were also conducted with the LibriSpeech ASR model, which is provided by the same source as the ASpIRE model. LibriSpeech ASR was trained on 960 hours from the LibriSpeech corpus, a collection of read English speech from public domain audio books. The model was abandoned in favor of the ASpIRE model due to its poor performance, stemming from the fact that read speech is much less varied and typically more formal than conversation speech. The discrepancy between LibriSpeech and ASpIRE performance demonstrates that formality in the training data is a significant determinant of model performance.

Per the LDC official page for the Fisher English corpus, "Under the Fisher protocol, a large number of participants each calls another participant, whom they typically do not know, for a short period of time to discuss the assigned topics. This maximizes inter-speaker variation and vocabulary breath while also increasing formality." (Fisher English Training Speech Part 1 Speech, n.d.). As such, CallHome's WER is not as low as the reported WER for Fisher English test set due to the differences in speaker demographics and relationship to each other. CallHome English speakers all contacted and spoke to someone they had a pre-existing, often close

relationship to, while Fisher English speakers were virtually strangers due to their differences in participant recruitment techniques.

Furthermore, CORAAL interviewers were generally close family friends or acquaintances of close family friends to the interviewees. CORAAL interviews often took place in the speaker's home or some other comfortable area in their local community. Additionally, interviewers were always African American themselves. All of these factors undoubtedly led to increased informality in the CORAAL conversations (Rickford and McNair-Knox, 1994). This increased informality may have encouraged more non-standard features to emerge in the interviewees' speech which would not have been present if they were instead participating in a widely organized telephone data collection project with predetermined topics, like that used for the Fisher English corpus.

Additionally, CallHome is used as a general representation of Northern, white American speech, but its speakers do not exclusively fit this persona. CallHome does include some speakers from other regions and a large portion of its speakers who received calls live abroad. There is a lack of rich background data on the CallHome speakers, much like the Fisher English corpus. Race and ethnic data are also not included, but, again, the manual audit notes from the LDC indicate that there is a presence of non-standard speech. The speakers marked as having Jewish or Yiddish accents were preserved in the data given to the ASpIRE model and altogether their average WER is 34.28. Without these speakers, the CallHome overall WER could have been as low as 21.71. Even still, the Jewish accented speakers are not a majority group in the CallHome data and as such, I believe the previous analysis from Section 5.1 and 5.2 holds. In fact, the difference in their inclusion or exclusion strengthens the argument that diverse representation in the training set is necessary in order to build a robust ASR system for all people.

CHAPTER 6

CONCLUSION

Given existing literature on racial disparities in ASR, this thesis set out to investigate the specific differences in model performance on speech from black speakers in CORAAL versus the speech in CallHome, a long-standing corpus used in many previous speech recognition projects. Four research questions were posed at the beginning of this thesis: (1) Does race affect ASR performance? (2) What specific features of AAL do ASR systems struggle with? (3) Do regional phonological patterns affect ASR performance? (4) Why does an ASR system perform differently on one type of speaker versus another? The first question prompted the study presented in this thesis and the following questions guided the experiments and analysis. They will be explicitly answered here.

6.1 Overall Trends Observed

Following racial divisions of the data sets, the ASpIRE model always performs worse given the CORAAL data. Thus, there is some effect of race on ASR performance, and, in this case, it is a negative effect.

Five common AAL phonological features were selected for further examination. Their analysis revealed that words which contained the phonological context suitable for a feature to appear were recognized at lower rates by the model for the CORAAL data, as seen in Table 1. The decreased performance of the model on these specific words versus the entire subset of common monosyllabic words shows that the features selected do all negatively affect model performance.

However, the negative affect is not consistent across the AAL features because model performance also follows regional associations. The ASpIRE model performs best on CallHome, then ROC - the two Northern or mostly Northern corpora. It then performs best on DCB, ATL, and lastly, PRV in terms of WER.

DCB and ATL are often at odds in terms of their WER ordering when examining the model performance in specific phonological contexts. In some cases (IY, AO, and all AAL features examined besides postvocalic word-final rhoticity), DCB outperforms ATL. In the others, ATL is better recognized by the model. This interchange is more than likely due to both cities' unique phonological standing due to their large non-native populations. Atlanta has been described as a Midland dialectal island in the South (Labov et al., 2006, pp. 261) and Washington D.C. does not fit squarely with any regional categorization but does show features from Southern and Northern speech. Nonetheless, some speakers in the ATL and DCB components did originate from other Southern states. As such, it follows that there are certainly some Southern speech patterns present in both datasets which would negatively affect model performance. The most Southern-representative dataset, PRV, is marked by the ASpIRE model's consistently poor performance.

6.2 Implications

Variation in AAL has had a complicated and hotly debated history, leading to the Uniformity Hypothesis and, subsequently, Homogeneity Myth (Labov, 1972; Wolfram and Kohn, 2015). The uniformity position purports there is a universality in AAL under the assumption that ethnicity invariably trumps regionality in regard to a speaker's dialect. Wolfram and Kohn (2015) remark this position as untrue and an oversimplification. In the results of this study, it is clear that model performance deteriorates given African American speakers. If all African Americans spoke AAL in the same manner, then these decreases in model performance would be analogous across

all the CORAAL data. The results do not demonstrate this pattern but reflect a likely additional role for regionality. The closer a speaker is to the regional (and racial) standard used in the training set, the better their speech is recognized.

It's possible that these staggered WERs in the AAL words reflect a quantitative difference in the AAL of different African American communities. One explanation is that there is a systematic variance in the frequency with which AAL phonological features appear. Alternatively, there could be qualitative distinction where the AAL features manifest differently between speakers of different cities. In general, it is imperative to recognize that Blackness and regionality interact in a significant manner, and these details cannot be taken in isolation of one another.

Thus, ROC slightly lags behind CallHome because despite its regional association, the speakers do still present AAL features in their speech which the Fisher English corpus does not capture sufficiently. Rochester speakers engage in AAL, but not in a Southern accented manner, which gives it a boost over the other CORAAL components in this case. Lack of Southern representation coupled with potential lack of African American representation in the training data primarily leaves Southern African American speakers at a larger disadvantage.

The farther a speaker's phoneme realizations stray from "the standard", the more difficult an ASR model will face when attempting to process the speech. As such, scrutiny is necessary when selecting training datasets for ASR systems to carefully define what the "standard" and target population are.

6.3 Recommendations for Future Work

Dialect specific models have been explored in the past to show isolated effects of more representative training data (Dorn, 2019). Significant improvements on Black speech are made when the training data of a model contains samples of Black speech to begin with. However, it

may not be feasible to employ multiple dialect-specific models for real world applications and use cases. Combining these smaller models with an ensemble method could be an optimal approach to limit training requirements while expanding diversity in the models' knowledge. Soto et al. (2016) presents one such study with shallow neural networks. A future project could train multiple smaller ASR models on AAL, Northern, and Southern dialects for an ensemble model.

An ensemble method such as that just mentioned may solve the errors sourced from lexicon deficiency. However, lexicon free ASR systems are also an interesting approach to ASR. Models using the Connectionist Temporal Classification (CTC) loss function can be character-based and eliminate the need for a word-based lexicon entirely (Graves et al., 2006; Graves and Jaitley, 2014). Future projects which train with AAL informed phone realizations in addition to standard phone realization could produce more robust ASR systems.

REFERENCES

Adams, C. A. (2009). *An acoustic phonetic analysis of African American English: A comparative study of two dialects*. Eastern Michigan University.

Audhkhasi, K., Ramabhadran, B., Saon, G., Picheny, M., & Nahamoo, D. (2017). *Direct Acoustics-to-Word Models for English Conversational Speech Recognition*.

Billa, J., Colhurst, T., El-Jaroudi, A., Iyer, R., Ma, K., Marsoukas, S., Quillen, C., Richardson, F., Siu, M., Zavaliagkos, G., & Gish, H. (1999). *Recent experiments in large vocabulary conversational speech recognition*. 41–44 vol.1. https://doi.org/10.1109/ICASSP.1999.758057

Britt, E., & Weldon, T. L. (2015). African American English in the Middle Class. In J. Bloomquist, L. J. Green, & S. L. Lanehart (Eds.), *The Oxford Handbook of African American Language*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199795390.013.44

Canavan, A., Graff, D., & Zipperlen, G. (1997). *CALLHOME American English Speech LDC97S42*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC97S42

Chapman, K. (2017). *The Northern Cities Shift: Minnesota's Ever-Changing Vowel Space* [Linguistics Honors Projects, Macalester College]. https://digitalcommons.macalester.edu/ling_honors/12

Cieri, C., Miller, D., & Walker, K. (2003). *From Switchboard to Fisher: Telephone Collection Protocols, Their Uses and Yields*. 4.

Cieri, C., Miller, D., & Walker, K. (2004). *The Fisher corpus: A resource for the next generations of speech-to-text*.

Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, *48*(6), 633–644. https://doi.org/10.1016/j.specom.2005.09.010

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, *118*(3), 1661–1676. https://doi.org/10.1121/1.2000774

Cohen Priva, U. (2017). Informativity and the Actuation of Lenition. *Language*, *93*(3), 569–597. https://doi.org/10.1353/lan.2017.0037

Cukor-Avila, P., & Balcazar, A. (2019). Exploring Grammatical Variation in the Corpus of Regional African American Language. *American Speech*, *94*(1), 36–53. https://doi.org/10.1215/00031283-7321989

Dorn, R. (2019). Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English. *Proceedings of the Student Research Workshop Associated with RANLP 2019*, 16–20. https://doi.org/10.26615/issn.2603-2821.2019_003

Elmahdy, M., Hasegawa-Johnson, M., & Mustafawi, E. (2013). *A Transfer Learning Approach for Under-Resourced Arabic Dialects Speech Recognition*. 5.

Erie Canal. (2021, January 15). Rochester Voices. http://www.rochestervoices.org/topics/erie-canal/

Farrington, C. (2018). Incomplete neutralization in African American English: The case of final consonant voicing. *Language Variation and Change*, *30*(3), 361–383. https://doi.org/10.1017/S0954394518000145

Farrington, C. (2020a, August). *AAL Linguistic Patterns*. Online Resources for African American Language. https://oraal.uoregon.edu/AAL/Linguistic-Patterns#Grammar_System

Farrington, C. (2020b, August 4). *What is AAL and who speaks it?* Online Resources for African American Language. https://oraal.uoregon.edu/AAL/What-is-AAL

Farrington, C. (2021, July 16). *CORAAL Components*. Online Resources for African American Language. https://oraal.uoregon.edu/coraal/components

Farrington, C., Kendall, T., Brooks, P. S., Jenson, L., Tacata, C., & Jaidan McLean. (2020). *The Corpus of Regional African American Language: ATL (Atlanta, GA 2017)* (2020.05) The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal

Farrington, C., & Schilling, N. (2019). Contextualizing The Corpus of Regional African American Language, D.C.:AAL in the Nation's Capital. *American Speech*, *94*(1), 21–35. https://doi.org/10.1215/00031283-7308060

Feagin, C. (2003). Vowel shifting in the southern states. In S. L. Sanders & S. J. Nagle (Eds.), *English in the Southern United States* (pp. 126–140). Cambridge University Press. https://doi.org/10.1017/CBO9780511486715.009

Fisher English Training Speech Part 1 Speech. (n.d.). Linguistic Data Consortium. Retrieved October 24, 2021, from https://catalog.ldc.upenn.edu/LDC2004S13

Forrest, J., & Wolfram, W. (2019). The Status Of (ING) in African American Language. *American Speech*, *94*(1), 72–90. https://doi.org/10.1215/00031283-7308049

Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., & Suendermann-Oeft, D. (2014). Comparing Open-Source Speech Recognition Toolkits. *NLPCS 2014: 11th International Workshop on Natural Language Processing and Cognitive Science*.

https://www.semanticscholar.org/paper/Comparing-Open-Source-Speech-Recognition-Toolkits-%E2%8B%86-Gaida-Lange/0c8fbe294172142b09d81a97b2e5b77113e08b42

Georgila, K., Leuski, A., Yanov, V., & Traum, D. (2020). Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6469–6476. https://aclanthology.org/2020.lrec-1.797

Graff, D., Dong, W., & Linguistic Data Consortium. (n.d.). *sph2pipe* (2.5) [Computer software]. https://www.ldc.upenn.edu/language-resources/tools/sphere-conversion-tools

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. https://doi.org/10.1145/1143844.1143891

Graves, A., & Jaitly, N. (2014). Towards End-to-End Speech Recognitionwith Recurrent Neural Networks. *Proceedings of the 31st International Conference on Machine Learning*, *32*, 1764–1772. https://proceedings.mlr.press/v32/graves14.html

Grieser, J. A. (2019). Investigating Topic-Based Style Shifting in the Classic Sociolinguistic Interview. *American Speech*, *94*(1), 54–71. https://doi.org/10.1215/00031283-7322011

Guglani, J., & Mishra, A. N. (2018). Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. *International Journal of Speech Technology*, *21*(2), 211–216. https://doi.org/10.1007/s10772-018-9497-6

Hermann, E., & Magimai-Doss, M. (2020). *Dysarthric Speech Recognition with Lattice-Free MMI*. 6109–6113. https://doi.org/10.1109/ICASSP40776.2020.9053549

Hinton, L. N., & Pollock, K. E. (2000). Regional Variations in the Phonological Characteristics of African American Vernacular English. *World Englishes*, *19*(1), 59–71. https://doi.org/10.1111/1467-971X.00155

Holliday, N. R. (2019). Variation in Question Intonation in the Corpus of Regional African American Language. *American Speech*, *94*(1), 110–130. https://doi.org/10.1215/00031283-7308038

Horton, W. S., & Gerrig, R. J. (2005). Conversational Common Ground and Memory Processes in Language Production. *Discourse Processes: A Multidisciplinary Journal*, *40*(1), 1–35.

Katerenchuk, D., Brizan, D. G., & Rosenberg, A. (2018, May). Interpersonal Relationship Labels for the CALLHOME Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1592

Kendall, T. (2019). New Perspectives on African American Language Through Public Corpora. *American Speech*, *94*(1), 13–20. https://doi.org/10.1215/00031283-7482427

Kendall, T. (2007). Enhancing Sociolinguistic Data Collections: The North Carolina Sociolinguistic Archive and Analysis Project. *Penn Working Papers in Linguistics 13.2*, 15–26.

Kendall, T., & Farrington, C. (2020). *The Corpus of Regional African American Language* (2020.05) The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal

Kendall, T., Quartey, M., Farrington, C., McLarty, J., Arnson, S., & Josley, Brooke. (2018). *The Corpus of Regional African American Language: DCB (Washington DC 2016)*

(2018.10.06) The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal

Khan, T. (2014). *Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic*. http://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-24645

King, S. (2018). *Exploring social and linguistic diversity across African Americans from Rochester, New York* [Ph. D dissertation, Stanford University]. https://searchworks.stanford.edu/view/12739840

King, S., Farrington, C., Kendall, T., Mullen, E., Arnson, S., & Jenson, L. (2020). *The Corpus of Regional African American Language: ROC (Rochester, NY 2016)* (2020.05) The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal

Kingsbury, P., Strassel, S., McLemore, C., & McIntyre, R. (1997). *CALLHOME American English Transcripts LDC97T14*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC97S42

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kohn, M. E. (2013). *Adolescent Ethnolinguistic Stability and Change: A Longitudinal Study* [Ph. D dissertation, University of North Carolina at Chapel Hill]. https://doi.org/10.17615/B1X3-P084

Kretzschmar, W. A. (2015). African American Voices in Atlanta. In J. Bloomquist, L. J. Green, & S. L. Lanehart (Eds.), *The Oxford Handbook of African American Language*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199795390.013.28

Kurata, G., Ramabhadran, B., Saon, G., & Sethy, A. (2017). *Language Modeling with Highway LSTM*. https://arxiv.org/abs/1709.06436

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. Univ. of Pennsylvania Press.

Labov, W. (1991). The three dialects of English. In P. Eckert (Ed.), *New Ways of Analyzing Sound Change* (pp. 1–44). Academic.

Labov, W. (2010). *Principles of linguistic change. 3: Cognitive and cultural factors* (1. publ). Wiley-Blackwell.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton.

Lehr, M., Gorman, K., & Shafran, I. (2014). Discriminative pronunciation modeling for dialectal speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1458–1462.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *In ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McLarty, J. (2020, June 26). *AAL Facts*. Online Resources for African American Language. https://oraal.uoregon.edu/facts#top

McLarty, J., Jones, T., & Hall, C. (2019). Corpus-Based Sociophonetic Approaches to Postvocalic R-Lessness in African American Language. *American Speech*, *94*(1), 91–109. https://doi.org/10.1215/00031283-7362239

Menon, R., Biswas, A., Saeb, A., Quinn, J., & Niesler, T. (2018). Automatic Speech Recognition for Humanitarian Applications in Somali. *ArXiv:1807.08669 [Cs, Stat]*. http://arxiv.org/abs/1807.08669

Millar, R. M. (2012). *English Historical Sociolinguistics*. Edinburgh University Press.

Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., & Traum, D. (2013). Which ASR should I choose for my dialogue system? *Proceedings of the SIGDIAL 2013 Conference*, 394–403. https://aclanthology.org/W13-4064

Povey, D. (n.d.). *Online decoding in Kaldi*. Kaldi. Retrieved October 19, 2021, from https://kaldi-asr.org/doc/online_decoding.html#online_decoding_nnet3

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011, December). The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Quartey, M., & Schilling, N. (2019). Shaping "Connected" versus "Disconnected" Identities in Narrative Discourse in D.C. African American Language. *American Speech*, *94*(1), 131–147. https://doi.org/10.1215/00031283-7322000

Renwick, M. E. L., & Olsen, R. M. (2017). Analyzing dialect variation in historical speech corpora. *The Journal of the Acoustical Society of America*, *142*(1), 406–421. https://doi.org/10.1121/1.4991009

Rickford, J. R. (1999). *African American vernacular English: Features, evolution, educational implications*. Blackwell Publishers.

Rickford, J. R., & McNair-Knox, F. (1994). Addressee- and Topic-Influenced Style Shift: A Quantitative Sociolinguistic Study. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (Vol. 1–xi, 385, pp. 235–276). Oxford University Press.

Rowe, R. (2005). *The Development of African American English in the Oldest Black Town in America: Plural -s Absence in Princeville, North Carolina* [North Carolina State University]. https://repository.lib.ncsu.edu/handle/1840.16/711

Rowe, R., Wolfram, Walt, Kendall, T., Farrington, C., & Josley, Brooke. (2018). *The Corpus of Regional African American Language: PRV (Princeville, NC 2004)* (2018.10.06) The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal

Sak, H., Senior, A., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *ArXiv:1402.1128 [Cs, Stat]*. http://arxiv.org/abs/1402.1128

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., & Hall, P. (2017). *English Conversational Telephone Speech Recognition by Humans and Machines*. http://arxiv.org/abs/1703.02136

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). *X-Vectors: Robust DNN Embeddings for Speaker Recognition*. 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Soto, V., Siohan, O., Elfeky, M., & Moreno, P. (2016). Selection and combination of hypotheses for dialectal speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5845–5849. https://doi.org/10.1109/ICASSP.2016.7472798

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Sue, D. W. (2013). Race talk: The psychology of racial dialogues. *American Psychologist*, *68*(8), 663–672. https://doi.org/10.1037/a0033681

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. https://doi.org/10.18653/v1/W17-1606

Tatman, R., & Kasten, C. (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Interspeech 2017*, 934–938. https://doi.org/10.21437/Interspeech.2017-1746

Thomas, E. R. (2007). Phonological and Phonetic Characteristics of African American Vernacular English: Phonological and Phonetic Characteristics of AAVE. *Language and Linguistics Compass*, *1*(5), 450–475. https://doi.org/10.1111/j.1749-818X.2007.00029.x

Thomas, E. R. (2008). Rural Southern White Accents. In E. W. Schneider & B. Kortmann (Eds.), *Varieties of English, 2: The Americas and the Caribbean* (Vol. 1–xxix, 800, pp. 87–114). Mouton de Gruyter.

U.S. Census Bureau. (2019). *Place of Birth by Nativity and Citizenship Status, 2019 American Community Survey 1-year Estimates*. https://data.census.gov/cedsci/table?text=B05002&g=0400000US11_1600000US1304000&tid=ACSDT1Y2019.B05002

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(3), 328–339. https://doi.org/10.1109/29.21701

Williams, J. D., Woodson, A. N., & Wallace, T. L. (2016). "Can We Say the N-word?": Exploring Psychological Safety During Race Talk. *Research in Human Development*, *13*(1), 15–31. https://doi.org/10.1080/15427609.2016.1141279

Wolfram, W., & Fasold, R. W. (1974). *The study of social dialects in American English*. Prentice-Hall.

Wolfram, W., & Kohn, M. E. (2015). Regionality in the Development of African American English. In J. Bloomquist, L. J. Green, & S. L. Lanehart (Eds.), *The Oxford Handbook of African American Language*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199795390.013.7

Yancy, G. (1995, September 8). Among Ourselves, the "N" Word Carries Warmth. *Philadelphia Tribune*, *112*(72), 6-A.

APPENDIX

Common Monosyllabic Words

The following table shows the list of the 333 monosyllabic words which appear at least ten times in each dataset. The table also presents the associated CMUdict pronunciation for a word, the average rate of correct identification by the model for that word (per subcorpus) and whether the word was considered to contain one of the AAL phonological contexts or regionally variable vowels of interest.

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
A	AH	0.7311	0.7025	0.6247	0.5632	0.5315	No	No
AGE	EY JH	1.0000	1.0000	0.7785	0.8421	0.5926	No	Yes
AH	AA	0.0347	0.0000	0.0000	0.0000	0.0000	No	No
AIN'T	EY N T	0.2143	0.1818	0.1127	0.2201	0.0952	Yes	Yes
ALL	AO L	0.8235	0.7756	0.7026	0.6076	0.5560	Yes	Yes
AM	AE M	0.7097	0.6327	0.5698	0.6486	0.5417	No	No
AN	AE N	0.6792	0.6630	0.5928	0.4783	0.5128	No	No
AND	AE N D	0.8130	0.7066	0.6224	0.6485	0.6054	Yes	No
ARE	AA R	0.6360	0.6288	0.5641	0.6897	0.3170	Yes	Yes
AS	AE Z	0.7289	0.8000	0.6439	0.6161	0.5479	No	No
ASK	AE S K	0.6667	0.4054	0.5532	0.4545	0.3800	No	No
AT	AE T	0.7049	0.6184	0.5245	0.5763	0.3626	No	No
BACK	B AE K	0.9342	0.8889	0.8008	0.7025	0.6687	No	No
BAD	B AE D	0.7965	0.8723	0.6538	0.6471	0.6852	No	No
BE	B IY	0.9092	0.8084	0.7511	0.6628	0.6984	No	Yes
BEEN	B IH N	0.9126	0.8346	0.7267	0.6732	0.5836	No	No
BEST	B EH S T	0.9333	0.6923	0.7419	0.7273	0.7273	Yes	No
BIG	B IH G	0.8672	0.8525	0.7560	0.7568	0.6161	No	No
BIT	B IH T	0.8876	0.8481	0.8821	0.7750	0.5946	No	No
BLACK	B L AE K	0.9200	0.8217	0.7505	0.6897	0.3235	No	No

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
BORN	B AO R N	1.0000	0.8302	0.6824	0.8571	0.4943	No	Yes
BOTH	B OW TH	0.8750	0.7805	0.7338	0.7826	0.4762	Yes	No
BOY	B OY	0.7667	0.8148	0.5400	0.4000	0.3333	No	No
BOYS	B OY Z	0.8182	0.7143	0.7688	0.3333	0.3056	No	No
BRING	B R IH NG	0.8750	0.7000	0.7059	0.7308	0.6389	No	No
BUS	B AH S	0.8667	0.8125	0.7067	0.7143	0.5676	No	Yes
BUT	B AH T	0.8113	0.6754	0.6204	0.6604	0.5416	No	Yes
BY	B AY	0.8435	0.7185	0.6005	0.7419	0.4167	No	Yes
CALL	K AO L	0.7724	0.6866	0.6087	0.7073	0.4907	Yes	Yes
CALLED	K AO L D	0.8770	0.5614	0.5308	0.3704	0.3714	Yes	Yes
CAME	K EY M	0.8898	0.8710	0.8546	0.8675	0.7500	No	Yes
CAN	K AE N	0.8766	0.7809	0.6151	0.6127	0.5875	No	No
CAN'T	K AE N T	0.8743	0.8741	0.7007	0.7182	0.6261	Yes	No
CAR	K AA R	0.8333	0.8780	0.6127	0.7857	0.5067	Yes	Yes
CARE	K EH R	0.9375	0.9167	0.7202	0.9545	0.5938	Yes	No
'CAUSE	K AH Z	0.5000	0.3216	0.3036	0.3435	0.1266	No	No
CHANGE	CH EY N JH	0.9643	0.9048	0.8000	0.4615	0.4348	No	Yes
CHECK	CH EH K	0.7600	0.8182	0.7917	0.6000	0.5000	No	No
CHILD	CH AY L D	0.7000	0.8182	0.7598	0.8000	0.5116	Yes	Yes
CLASS	K L AE S	0.9167	0.8571	0.8631	0.8810	0.5588	No	No
CLOSE	K L OW S	0.9310	0.9032	0.8098	0.8519	0.7606	No	No
COME	K AH M	0.8571	0.8444	0.7599	0.6644	0.7093	No	Yes
COMES	K AH M Z	0.8529	0.9200	0.7705	0.9565	0.6875	No	Yes
COOL	K UW L	0.8511	0.5252	0.6943	0.7169	0.2143	Yes	No
COULD	K UH D	0.8025	0.6215	0.4781	0.4403	0.5109	No	No
COURSE	K AO R S	0.9070	0.8214	0.7297	0.8621	0.4000	No	Yes
DAD	D AE D	0.6415	0.7083	0.6621	0.7143	0.4556	No	No
DAY	D EY	0.8792	0.8280	0.7675	0.5526	0.6150	No	Yes
DAYS	D EY Z	0.9083	0.9444	0.7640	0.7619	0.5238	No	Yes
DEAL	D IY L	0.9394	0.8333	0.7805	0.5385	0.5882	Yes	Yes
DID	D IH D	0.7535	0.7266	0.5911	0.5515	0.3919	No	No
DO	D UW	0.7773	0.7919	0.7043	0.7361	0.5686	No	No
DOES	D AH Z	0.7265	0.7045	0.4927	0.6667	0.1724	No	Yes
DOING	D UW IH NG	0.8969	0.8671	0.6725	0.7528	0.5828	No	No
DONE	D AH N	0.8615	0.5738	0.5142	0.2308	0.4595	No	Yes

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
DON'T	D OW N T	0.8802	0.8074	0.6992	0.6404	0.6273	Yes	No
DOWN	D AW N	0.8836	0.7875	0.7535	0.7520	0.6177	No	No
DRIVE	D R AY V	0.8462	0.8571	0.6154	0.7000	0.4857	No	Yes
EACH	IY CH	0.9487	0.9667	0.8512	0.9565	0.6889	No	Yes
EAT	IY T	0.7714	0.7692	0.5414	0.7333	0.6071	No	Yes
EIGHT	EY T	0.8261	0.7667	0.6982	0.7727	0.6267	No	Yes
ELSE	EH L S	0.8659	0.8780	0.7261	0.6750	0.5352	Yes	No
END	EH N D	0.8542	0.6400	0.6139	0.5238	0.3514	Yes	No
FACT	F AE K T	0.8889	1.0000	0.8500	0.5484	0.5000	Yes	No
FAR	F AA R	0.8302	0.9600	0.7639	0.7581	0.5397	Yes	Yes
FEEL	F IY L	0.8790	0.7542	0.6938	0.2681	0.7674	Yes	Yes
FELT	F EH L T	0.8205	0.9020	0.6057	0.8000	0.5333	Yes	No
FEW	F Y UW	0.8776	0.8431	0.8556	0.7727	0.6346	No	No
FIND	F AY N D	0.9000	1.0000	0.7518	0.8125	0.5333	Yes	Yes
FIRST	F ER S T	0.9745	0.9328	0.8663	0.8495	0.7553	Yes	No
FIVE	F AY V	0.9545	0.9444	0.7411	0.7465	0.6161	No	Yes
FOOD	F UW D	0.9130	0.9259	0.7500	0.8696	0.6087	No	No
FOR	F AO R	0.8822	0.8647	0.7820	0.6674	0.6371	Yes	Yes
FOUR	F AO R	0.7228	0.8800	0.6745	0.6286	0.4510	Yes	Yes
FREE	F R IY	0.8000	0.7826	0.7917	0.8333	0.9000	No	Yes
FRIEND	F R EH N D	0.8723	0.9032	0.7569	0.7778	0.5000	Yes	No
FROM	F R AH M	0.9136	0.9065	0.8375	0.8317	0.6229	No	Yes
FRONT	F R AH N T	0.9167	0.9167	0.8588	0.9286	0.6552	Yes	Yes
FUN	F AH N	0.8788	0.8182	0.6904	0.8235	0.6296	No	Yes
GET	G EH T	0.9086	0.8698	0.7846	0.7564	0.6743	No	No
GIRL	G ER L	0.8049	0.7959	0.6010	0.5862	0.5600	Yes	No
GIVE	G IH V	0.8706	0.8857	0.5975	0.6750	0.5000	No	No
GO	G OW	0.8714	0.8893	0.6943	0.7645	0.7154	No	No
GOD	G AA D	0.6746	0.7568	0.4956	0.4375	0.5696	No	Yes
GOOD	G UH D	0.8448	0.8249	0.7516	0.7184	0.5539	No	No
GOT	G AA T	0.8205	0.7742	0.7029	0.6725	0.6410	No	Yes
GRADE	G R EY D	0.5833	0.7257	0.6509	0.5833	0.7736	No	Yes
GREAT	G R EY T	0.8321	0.6800	0.7565	0.6957	0.5441	No	Yes
GUESS	G EH S	0.9481	0.8212	0.8175	0.8765	0.7857	No	No
GUY	G AY	0.8393	0.8000	0.7455	0.6129	0.5455	No	Yes

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
GUYS	G AY Z	0.9167	0.8095	0.8235	0.8462	0.4348	No	Yes
HA	HH AA	0.0000	0.0000	0.0000	0.0000	0.0000	No	No
HAD	HH AE D	0.8240	0.7567	0.6594	0.6640	0.5423	No	No
HALF	HH AE F	0.9082	0.8750	0.6301	0.8000	0.5833	No	No
HANG	HH AE NG	0.9091	0.8182	0.7941	0.8333	0.6400	No	No
HARD	HH AA R D	0.8684	0.8409	0.7227	0.6047	0.5313	Yes	Yes
HAS	HH AE Z	0.6712	0.7452	0.5956	0.7609	0.5000	No	No
HATE	HH EY T	0.9545	0.7857	0.6875	0.9000	0.4000	No	Yes
HAVE	HH AE V	0.8506	0.8325	0.7413	0.7415	0.6264	No	No
HE	HH IY	0.7309	0.7348	0.6100	0.5568	0.5033	No	Yes
HEAD	HH EH D	0.6800	0.5000	0.4844	0.5714	0.1714	No	No
HEAR	HH IY R	0.8514	0.6061	0.4696	0.4419	0.4237	Yes	Yes
HEARD	HH ER D	0.8632	0.7447	0.6748	0.5400	0.6098	No	No
HEART	HH AA R T	0.8519	0.7000	0.5185	0.5714	0.5714	Yes	Yes
HELP	HH EH L P	0.9268	0.7179	0.6405	0.6522	0.3277	Yes	No
HER	HH ER	0.7342	0.7317	0.5615	0.5147	0.3856	Yes	No
HERE	HH IY R	0.8609	0.8390	0.6704	0.6923	0.5616	Yes	Yes
HE'S	HH IY Z	0.7970	0.6709	0.6330	0.5979	0.4819	No	Yes
HEY	HH EY	0.7241	0.6774	0.4556	0.5161	0.4688	No	Yes
HIGH	HH AY	0.7805	0.8492	0.8282	0.6753	0.5164	No	Yes
HIM	HH IH M	0.6140	0.5683	0.5111	0.4667	0.4208	No	No
HIS	HH IH Z	0.7686	0.7829	0.6561	0.5603	0.5300	No	No
HM	HH AH M	0.0000	0.0000	0.0000	0.0000	0.0000	No	No
HOME	HH OW M	0.8816	0.9182	0.8268	0.8710	0.7811	No	No
HOPE	HH OW P	0.7381	0.6500	0.6818	0.2727	0.4375	No	No
HOT	HH AA T	0.6250	0.7273	0.4828	0.7857	0.6875	No	Yes
HOUSE	HH AW S	0.8718	0.9254	0.7962	0.8000	0.6486	No	No
HOW	HH AW	0.7813	0.8537	0.7315	0.6963	0.5667	No	No
HUH	HH AH	0.2218	0.0769	0.1200	0.4000	0.0513	No	No
I	AY	0.8182	0.7626	0.7103	0.6683	0.5945	No	Yes
I'D	AY D	0.5698	0.4706	0.3679	0.3889	0.1304	No	Yes
IF	IH F	0.7886	0.6728	0.5752	0.5758	0.4492	No	No
I'LL	AY L	0.6205	0.5286	0.3317	0.2000	0.2258	Yes	Yes
I'M	AY M	0.7362	0.6507	0.6066	0.5584	0.4569	No	Yes
IN	IH N	0.7381	0.6654	0.6051	0.6018	0.4866	No	No

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
IS	IH Z	0.6576	0.6391	0.5428	0.5575	0.4023	No	No
IT	IH T	0.7320	0.6591	0.5626	0.5396	0.4500	No	No
IT'S	IH T S	0.7809	0.5839	0.4419	0.4532	0.2988	No	No
I'VE	AY V	0.7531	0.7339	0.6032	0.5435	0.5208	No	Yes
JOB	JH AA B	0.9518	0.9885	0.9231	0.7667	0.7234	No	Yes
JUST	JH AH S T	0.8861	0.7958	0.7141	0.7051	0.5706	Yes	Yes
KEEP	K IY P	0.9592	0.9524	0.8421	0.8478	0.6667	No	Yes
KEPT	K EH P T	0.7273	0.8571	0.7179	0.5385	0.5641	Yes	No
KID	K IH D	0.7000	0.5957	0.6091	0.5000	0.4815	No	No
KIDS	K IH D Z	0.9063	0.8254	0.7703	0.8684	0.6463	No	No
KIND	K AY N D	0.7684	0.7390	0.6553	0.6746	0.5338	Yes	Yes
KNEW	N UW	0.7414	0.6792	0.7428	0.5714	0.6000	No	No
KNOW	N OW	0.8702	0.7602	0.7477	0.7603	0.6292	No	No
LAST	L AE S T	0.8929	0.8167	0.8198	0.8000	0.6667	Yes	No
LEARN	L ER N	0.8889	0.8400	0.7308	0.8947	0.5625	No	No
LEAST	L IY S T	0.8333	0.9091	0.7375	0.5172	0.5333	Yes	Yes
LEFT	L EH F T	0.8906	0.7632	0.6667	0.4706	0.3750	Yes	No
LET	L EH T	0.7500	0.8056	0.6257	0.4545	0.4696	No	No
LET'S	L EH T S	0.7021	0.4894	0.4900	0.5887	0.3721	No	No
LIFE	L AY F	0.9333	0.9103	0.7849	0.8088	0.6500	No	Yes
LIKE	L AY K	0.8977	0.7913	0.7341	0.7250	0.6178	No	Yes
LINE	L AY N	0.8000	0.7857	0.6207	0.3333	0.6875	No	Yes
LIVE	L IH V	0.8806	0.8739	0.7196	0.6897	0.6094	No	No
LIVED	L IH V D	0.5714	0.6514	0.5796	0.6316	0.3922	Yes	No
LONG	L AO NG	0.8241	0.9091	0.7888	0.7069	0.7042	No	Yes
LOOK	L UH K	0.7980	0.7048	0.6611	0.5888	0.6078	No	No
LOOKED	L UH K T	0.7714	0.4545	0.4787	0.4167	0.2368	No	No
LOST	L AO S T	0.7500	0.8636	0.7206	0.6923	0.6667	Yes	Yes
LOT	L AA T	0.8813	0.8358	0.7926	0.7463	0.6737	No	No
LOVE	L AH V	0.8545	0.7864	0.6923	0.7193	0.5426	No	Yes
MADE	M EY D	0.8625	0.8421	0.6702	0.6579	0.5595	No	Yes
MAKE	M EY K	0.8712	0.9028	0.7610	0.7442	0.6560	No	Yes
MAKES	M EY K S	0.8056	0.9032	0.7404	0.5625	0.5714	No	Yes
MAN	M AE N	0.6531	0.6724	0.4439	0.4670	0.5481	No	No
MAY	M EY	0.6098	0.7500	0.6693	0.8000	0.5135	No	Yes

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
ME	M IY	0.8807	0.8629	0.7503	0.4885	0.6159	No	Yes
MEAN	M IY N	0.8500	0.7700	0.6959	0.8041	0.5516	No	Yes
MET	M EH T	0.9444	0.8158	0.7067	0.5000	0.3333	No	No
MIGHT	M AY T	0.8642	0.8542	0.6779	0.5773	0.6140	No	Yes
MIND	M AY N D	0.7600	0.7200	0.6116	0.6579	0.5000	Yes	Yes
MM	EH M	0.1880	0.0526	0.0879	0.2241	0.2019	No	No
MOM	M AA M	0.7917	0.7778	0.7476	0.8182	0.6708	No	Yes
MONTHS	M AH N TH S	0.9565	0.7895	0.8356	0.5833	0.6000	Yes	Yes
MORE	M AO R	0.9202	0.9059	0.8120	0.8429	0.5753	Yes	Yes
MOST	M OW S T	0.8864	0.9459	0.8456	0.7805	0.7983	Yes	No
MOVE	M UW V	0.8667	0.7755	0.6667	0.6667	0.6232	No	No
MOVED	M UW V D	0.8750	0.7722	0.7493	0.7391	0.5289	Yes	No
MUCH	M AH CH	0.9268	0.9559	0.8927	0.8504	0.7212	No	Yes
MUSIC	M Y UW Z IH K	0.9333	0.8333	0.7082	0.7500	0.4545	No	No
MY	M AY	0.8842	0.8687	0.7899	0.7523	0.6986	No	Yes
NAME	N EY M	0.8545	0.8148	0.6450	0.5875	0.5600	No	Yes
NEED	N IY D	0.7907	0.7297	0.5965	0.5946	0.5862	No	Yes
NEW	N UW	0.8851	0.8333	0.7821	0.7935	0.7935	No	No
NEXT	N EH K S T	0.9252	0.9683	0.7800	0.7778	0.5714	Yes	No
NICE	N AY S	0.9140	0.8689	0.8418	0.7500	0.4809	No	Yes
NIGHT	N AY T	0.8286	0.9200	0.6432	0.8462	0.6667	No	Yes
NINE	N AY N	0.9556	0.9630	0.5487	0.8462	0.5738	No	Yes
NO	N OW	0.7467	0.6154	0.5263	0.5425	0.4390	No	No
NOT	N AA T	0.8625	0.8354	0.6908	0.7157	0.6716	No	Yes
NOW	N AW	0.8186	0.8013	0.7179	0.6779	0.4633	No	No
OF	AH V	0.8238	0.7720	0.7079	0.6680	0.5551	No	Yes
OFF	AO F	0.8716	0.7500	0.5490	0.5775	0.4200	No	Yes
OH	OW	0.6638	0.6361	0.4350	0.4568	0.2449	No	No
OLD	OW L D	0.7470	0.7534	0.5207	0.5957	0.3238	Yes	No
ON	AA N	0.8220	0.7462	0.6364	0.6448	0.5689	No	No
ONCE	W AH N S	0.7349	0.7755	0.6163	0.8077	0.6471	No	Yes
ONE	W AH N	0.8853	0.8139	0.7597	0.6875	0.6809	No	Yes
ONES	W AH N Z	0.8261	0.7222	0.6563	0.4762	0.5385	No	Yes
OR	AO R	0.7581	0.6076	0.4739	0.5932	0.3133	Yes	Yes
OUR	AW ER	0.6975	0.6561	0.4839	0.5439	0.2555	Yes	No

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
OUT	AW T	0.8290	0.7960	0.7062	0.6530	0.5469	No	No
OWN	OW N	0.9184	0.8772	0.7734	0.8077	0.6353	No	No
PART	P AA R T	0.8537	0.8254	0.7575	0.7368	0.5357	Yes	Yes
PAY	P EY	0.9455	0.8108	0.8148	0.8333	0.6000	No	Yes
PHONE	F OW N	0.7647	0.7857	0.8226	0.6250	0.7959	No	No
PICK	P IH K	1.0000	0.7500	0.6825	0.6923	0.7273	No	No
PLACE	P L EY S	0.9130	0.8837	0.7440	0.8919	0.6625	No	Yes
PLAY	P L EY	0.8108	0.7826	0.7508	0.6279	0.6567	No	Yes
PLAYED	P L EY D	0.9091	0.5714	0.3391	0.3000	0.1481	No	Yes
POINT	P OY N T	0.9756	0.8750	0.7740	0.8444	0.4310	Yes	No
PUT	P UH T	0.9320	0.8310	0.7380	0.7412	0.5655	No	No
REAL	R IY L	0.8837	0.8696	0.6384	0.4074	0.7027	Yes	Yes
RIGHT	R AY T	0.8529	0.7022	0.6496	0.7162	0.5863	No	Yes
RUN	R AH N	1.0000	0.8235	0.6316	0.5455	0.5094	No	Yes
SAID	S EH D	0.7820	0.6040	0.4679	0.4216	0.3340	No	No
SAME	S EY M	0.8923	0.9416	0.8176	0.8871	0.5778	No	Yes
SAW	S AO	0.7500	0.8000	0.5855	0.6364	0.4375	No	Yes
SAY	S EY	0.7749	0.7979	0.6997	0.7939	0.5142	No	Yes
SAYS	S EH Z	0.7162	0.8462	0.6200	0.7000	0.3871	No	No
SCHOOL	S K UW L	0.9510	0.8854	0.8499	0.8800	0.6865	Yes	No
SEE	S IY	0.8504	0.8553	0.7425	0.7539	0.6515	No	Yes
SEEM	S IY M	0.6923	1.0000	0.6667	0.3333	0.2353	No	Yes
SEEN	S IY N	0.9038	0.7742	0.7251	0.4595	0.4630	No	Yes
SENSE	S EH N S	0.9444	0.7436	0.6941	0.5238	0.6000	No	No
SET	S EH T	0.7632	0.7037	0.4839	0.6429	0.5172	No	No
SHE	SH IY	0.8447	0.8852	0.7656	0.7014	0.6723	No	Yes
SHE'S	SH IY Z	0.8079	0.7921	0.7396	0.7619	0.5977	No	Yes
SHOULD	SH UH D	0.8252	0.7059	0.6790	0.7222	0.6190	No	No
SHOW	SH OW	0.8000	0.8108	0.6737	0.8077	0.5714	No	No
SIDE	S AY D	0.7826	0.8542	0.4907	0.8077	0.3472	No	Yes
SINCE	S IH N S	0.8519	0.8788	0.5872	0.8000	0.4677	No	No
SIT	S IH T	0.8571	0.7586	0.5612	0.6364	0.5476	No	No
SIX	S IH K S	0.9041	0.9429	0.8026	0.9394	0.5278	No	No
SO	S OW	0.8022	0.8471	0.7876	0.8058	0.6198	No	No
SOME	S AH M	0.8482	0.8421	0.7545	0.6156	0.5928	No	Yes

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
SOUND	S AW N D	0.8571	0.4750	0.4394	0.4643	0.5455	Yes	No
STAND	S T AE N D	0.9000	0.9286	0.5870	0.5000	0.4615	Yes	No
START	S T AA R T	0.8542	0.6346	0.5436	0.6383	0.4176	Yes	Yes
STATE	S T EY T	0.7143	0.6875	0.5577	0.6875	0.4595	No	Yes
STAY	S T EY	0.8065	0.7586	0.6624	0.9118	0.5276	No	Yes
STAYED	S T EY D	0.6250	0.5833	0.5313	0.5000	0.3217	No	Yes
STILL	S T IH L	0.8596	0.8039	0.6999	0.6134	0.5238	Yes	No
STOP	S T AA P	0.5294	0.6667	0.5294	0.6429	0.6857	No	Yes
STORE	S T AO R	0.6000	0.7083	0.7895	0.6111	0.4624	Yes	Yes
STREET	S T R IY T	0.9412	0.7453	0.6764	0.6452	0.6040	No	Yes
STUFF	S T AH F	0.9022	0.8543	0.7732	0.7882	0.6491	No	Yes
SUCH	S AH CH	0.8772	0.9583	0.7255	0.9231	0.3750	No	Yes
SURE	SH UH R	0.9231	0.8571	0.7116	0.6000	0.4345	No	No
TAKE	T EY K	0.9282	0.8923	0.7438	0.6667	0.6716	No	Yes
TALK	T AO K	0.8652	0.7500	0.6931	0.4857	0.6196	No	Yes
TAUGHT	T AO T	0.8571	0.9286	0.5046	0.7500	0.5000	No	Yes
TEACH	T IY CH	0.8000	0.9091	0.7778	0.8333	0.7000	No	Yes
TELL	T EH L	0.8683	0.6712	0.6763	0.5139	0.4574	Yes	No
TEN	T EH N	0.8621	0.6977	0.6434	0.7083	0.6000	No	No
THAN	DH AE N	0.7549	0.6827	0.5930	0.4725	0.3305	Yes	No
THANK	TH AE NG K	0.7778	0.3636	0.4268	0.4286	0.3235	Yes	No
THAT	DH AE T	0.8303	0.8078	0.6950	0.5504	0.5281	Yes	No
THAT'S	DH AE T S	0.7758	0.6906	0.6290	0.5395	0.4107	Yes	No
THE	DH AH	0.8398	0.7560	0.6813	0.6256	0.5415	Yes	No
THEIR	DH EH R	0.7655	0.7017	0.5891	0.6600	0.4337	Yes	No
THEM	DH EH M	0.7737	0.8319	0.6605	0.5440	0.5968	Yes	No
THEN	DH EH N	0.7940	0.7163	0.5976	0.5439	0.4433	Yes	No
THERE	DH EH R	0.8046	0.7708	0.6229	0.6507	0.4665	Yes	No
THERE'S	DH EH R Z	0.7317	0.6196	0.5850	0.6170	0.2381	Yes	No
THESE	DH IY Z	0.8936	0.8487	0.7281	0.5862	0.6026	Yes	Yes
THEY	DH EY	0.8850	0.7862	0.7132	0.5929	0.6008	Yes	Yes
THEY'LL	DH EH L	0.5806	0.4706	0.2353	0.3810	0.2222	Yes	No
THEY'RE	DH EH R	0.6565	0.5733	0.5280	0.6769	0.3364	Yes	No
THING	TH IH NG	0.8588	0.8095	0.7365	0.7797	0.6067	Yes	No
THINGS	TH IH NG Z	0.9159	0.8953	0.8410	0.8587	0.7108	Yes	No

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
THINK	TH IH NG K	0.9288	0.8799	0.7955	0.6326	0.5563	Yes	No
THIS	DH IH S	0.8024	0.7056	0.5834	0.5199	0.5824	Yes	No
THOSE	DH OW Z	0.8542	0.8198	0.8246	0.8167	0.5424	Yes	No
THOUGH	DH OW	0.7000	0.6951	0.4810	0.2932	0.4375	Yes	No
THOUGHT	TH AO T	0.8075	0.7973	0.6637	0.4667	0.5254	Yes	Yes
THREE	TH R IY	0.9588	0.9221	0.7920	0.7742	0.7213	Yes	Yes
THROUGH	TH R UW	0.8500	0.6667	0.6174	0.3889	0.5159	Yes	No
TIME	T AY M	0.8889	0.9046	0.7902	0.8075	0.6540	No	Yes
TIMES	T AY M Z	0.9286	0.7838	0.6871	0.7500	0.3913	No	Yes
TO	T UW	0.8393	0.8125	0.7266	0.6902	0.5710	No	No
TOLD	T OW L D	0.8834	0.7816	0.6516	0.5000	0.5537	Yes	No
TOO	T UW	0.7660	0.7483	0.6493	0.5170	0.5245	No	No
TOOK	T UH K	0.9259	0.9048	0.7265	0.4643	0.5057	No	No
TOP	T AA P	0.8182	0.8889	0.7317	0.6250	0.3784	No	No
TOUCH	T AH CH	0.9474	0.8000	0.4706	0.6000	0.7692	No	Yes
TOWN	T AW N	1.0000	0.7887	0.6806	0.7000	0.6298	No	No
TRY	T R AY	0.8727	0.8615	0.6786	0.5800	0.7595	No	Yes
TURN	T ER N	0.9286	0.4667	0.5185	0.4286	0.5200	No	No
TWO	T UW	0.7891	0.7708	0.6848	0.6341	0.5819	No	No
TYPE	T AY P	0.8750	0.9800	0.7630	0.7770	0.6500	No	Yes
UH	AH	0.3009	0.5497	0.4219	0.4659	0.3128	No	Yes
UM	AH M	0.8208	0.5237	0.5453	0.6492	0.6077	No	Yes
UP	AH P	0.8448	0.8395	0.6735	0.5605	0.5007	No	Yes
US	AH S	0.8601	0.7619	0.6534	0.7015	0.4690	No	Yes
USE	Y UW Z	0.8704	0.7600	0.7603	0.7241	0.6111	No	No
USED	Y UW Z D	0.7966	0.7283	0.6236	0.6623	0.5292	Yes	No
WALK	W AO K	0.8400	0.7000	0.6931	0.7826	0.4571	No	No
WALKED	W AO K T	0.7000	0.4737	0.3951	0.1818	0.1250	Yes	Yes
WANT	W AA N T	0.8031	0.7847	0.6730	0.7361	0.4224	Yes	No
WAS	W AA Z	0.8292	0.7768	0.7272	0.6963	0.5621	No	Yes
WATCH	W AA CH	0.8276	0.9000	0.7570	0.6857	0.6000	No	No
WAY	W EY	0.8471	0.8187	0.7675	0.7436	0.6444	No	Yes
WE	W IY	0.8645	0.8120	0.7168	0.6529	0.6255	No	Yes
WEEK	W IY K	0.9167	0.8286	0.7660	0.5000	0.6744	No	Yes
WELL	W EH L	0.7209	0.5602	0.4908	0.5500	0.2713	Yes	No

		Average % Correct					Contains	
Word	CMUdict Pronunciation	Call Home	ROC	DCB	ATL	PRV	AAL Word?	Regional Vowel?
WE'LL	W IH L	0.5429	0.3500	0.2845	0.2500	0.1613	Yes	No
WENT	W EH N T	0.8750	0.7465	0.6924	0.6593	0.5127	Yes	No
WERE	W ER	0.7011	0.6032	0.5824	0.6197	0.4198	Yes	No
WE'RE	W IY R	0.7299	0.5843	0.4917	0.5000	0.3038	Yes	No
WHAT	HH W AH T	0.7971	0.7430	0.6253	0.5496	0.4709	No	Yes
WHAT'S	HH W AH T S	0.8362	0.6623	0.4671	0.4326	0.2826	No	Yes
WHEN	HH W EH N	0.8277	0.7748	0.6880	0.7104	0.5164	No	No
WHERE	HH W EH R	0.7614	0.7642	0.5756	0.6309	0.3795	No	No
WHICH	HH W IH CH	0.9138	0.8174	0.7428	0.9063	0.5326	No	No
WHILE	HH W AY L	0.8295	0.7500	0.5214	0.6087	0.4242	Yes	Yes
WHITE	HH W AY T	0.8095	0.7634	0.6775	0.5424	0.4615	No	Yes
WHO	HH UW	0.7212	0.7177	0.6273	0.6029	0.3486	No	No
WHOLE	HH OW L	0.9109	0.8955	0.8457	0.8182	0.7071	Yes	No
WHY	HH W AY	0.6883	0.8167	0.5272	0.5818	0.5116	No	Yes
WILL	W IH L	0.6131	0.5750	0.5345	0.5962	0.3182	Yes	No
WITH	W IH DH	0.8365	0.7755	0.6372	0.5049	0.5480	Yes	No
WORD	W ER D	0.8000	0.2727	0.4659	0.5000	0.1818	No	No
WORK	W ER K	0.9293	0.8371	0.7333	0.6579	0.4569	No	No
WORLD	W ER L D	0.8649	0.6786	0.6809	0.7568	0.3684	Yes	No
WOULD	W UH D	0.7681	0.6743	0.5153	0.5526	0.4253	No	No
WOW	W AW	0.8598	0.5091	0.4967	0.5147	0.3952	No	No
WRONG	R AO NG	0.9048	0.8095	0.7798	0.9091	0.5652	No	Yes
YEAH	Y AE	0.8675	0.6914	0.5668	0.5837	0.5756	No	No
YEAR	Y IH R	0.8742	0.7642	0.6564	0.7872	0.5242	Yes	No
YEARS	Y IH R Z	0.9464	0.9816	0.7880	0.8831	0.7014	No	No
YES	Y EH S	0.7339	0.5278	0.5241	0.5745	0.5205	No	No
YET	Y EH T	0.7093	0.5217	0.4789	0.4667	0.2353	No	No
YORK	Y AO R K	0.9778	0.8191	0.9176	0.9048	0.9000	No	Yes
YOU	Y UW	0.8318	0.7338	0.6973	0.6476	0.5623	No	No
YOUNG	Y AH NG	0.9231	0.8800	0.8445	0.8125	0.7059	No	Yes
YOUR	Y AO R	0.7422	0.7313	0.6728	0.5921	0.4721	Yes	Yes
YOU'RE	Y UH R	0.7095	0.6488	0.5817	0.6623	0.2764	Yes	No
YUP	Y AH P	0.3571	0.2558	0.2038	0.3704	0.0930	No	No