ANALYSIS OF TWO ACOUSTIC MODELS ON FORCED ALIGNMENT OF AFRICAN AMERICAN ENGLISH

by

SIERRA MAGNOTTA

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

ABSTRACT

Automatic speech recognition (ASR) is the process by which spoken speech is recognized and transcribed by a system. Forced alignment is a task within speech recognition that outputs a time-aligned transcript of audio at the phoneme and word level, utilizing an acoustic model that represents the relationship between audio signal and linguistic phonemes. This thesis compares two acoustic models, one trained on African American English (AAE) varieties and one trained on Mainstream U.S. English, on a forced alignment task of AAE speakers from Georgia. The output from each system's forced alignment was analyzed to find differences in performance between the two acoustic models. We find that the two systems differ significantly in reported vowel duration for certain vowels relevant to ongoing changes in AAE.

INDEX WORDS: Automatic Speech Recognition, Forced Alignment, Acoustic Model, African American Language, Computational Linguistics

ANALYSIS OF TWO ACOUSTIC MODELS ON FORCED ALIGNMENT OF AFRICAN AMERICAN ENGLISH

by

SIERRA MAGNOTTA

B.A., Bucknell University, 2018

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

2022

ANALYSIS OF TWO ACOUSTIC MODELS ON FORCED ALIGNMENT OF AFRICAN AMERICAN ENGLISH

by

SIERRA MAGNOTTA

Major Professor:	Margaret E. L. Renwick
Committee:	John Hale
	Jon Forrest

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2022

ACKNOWLEDGEMENTS

I would like to thank Dr. Renwick for her support through this entire process. I am extremely grateful for her understanding and guidance. I also would like to thank Dr. Hale and Dr. Forrest for their assistance and for serving on my committee. Thank you to Joey Stanley for his help with writing Praat scripts and his wonderful tutorials. Lastly, thank you to my friends and family who believed in me and supported me along the way.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background

African American Language (AAL) is a well-studied dialect of English that has its own grammatical system, phonology, and lexical items and is spoken by most Black people in the United States (Kendall et al. 2018). Automatic speech recognition (ASR) is an area within computational linguistics and artificial intelligence where computer systems are developed to parse spoken language and output the speech as text. While ASR systems have made huge progress in the last decade, there are still many unsolved issues in the field. Specifically, ASR systems are influenced by the training data they are given to first learn to recognize speech (Koenecke et al. 2020). The training data typically lack diversity in the speakers and varieties of language used, leading to performance disparities.

This thesis builds on the current research in this area by analyzing the impact of training data used on an ASR system's acoustic model. The acoustic model represents the relationship between the spoken audio signal and the linguistic phonemes being produced in the audio, and research suggests that the data used for training the acoustic model may be responsible for performance disparities in ASR systems (Koenecke et al. 2020). This thesis compares the performance of two speech recognition systems that are identical except for the acoustic model. One system used an acoustic model trained on a general variety of American English, and one system used an acoustic model trained on speech from African American speakers collected through CORAAL (Kendall and Farrington 2021). Both systems were used on a forced alignment

task of African American speech data from two locations in Georgia. Forced alignment is a different task than automatic speech recognition; given audio and a sentence-level transcript of the audio, forced aligners provide word and phoneme-level transcriptions of the audio. The results of the two acoustic model systems on this forced alignment task show broad agreement in representations of AAL consonantal features, but less agreement in some vowel contexts, especially relating to vowel duration. Differences in vowel durations and onsets are shown to be statistically significant in certain environments.

Research has shown that ASR systems perform poorly on speech data that is not well-represented in the system's training data, which typically consists of predominantly white Mainstream United States English (MUSE) speakers. Koenecke et al. examined this by testing five commercial ASR systems - Amazon, Apple, Google, IBM, and Microsoft - on different varieties of English and found that all the reported significantly higher word error rates (WER) on Black speakers (2020). Additionally, their results showed that WER was strongly associated with the amount of AAL features being used by the speaker – the more AAL features a speaker used, the worse the systems' resulting WER was (Koenecke et al. 2020). Tatman and Kasten (2017) studied differences in gender, race, and regional dialects for both Bing Speech and YouTube automatic captions. Their results showed that YouTube captions performed significantly worse on Black speakers (Tatman and Kasten 2017). Bing Speech also performed worse on Black speakers, but the sample size was too small to achieve high power (Tatman and Kasten 2017).

These and similar studies show that ASR does indeed perform worse on AAL, but the reasons for this are contested. Some research has shown that ASR systems are more likely to produce errors around morpho-syntactic features of AAL. Martin and Tang (2020) examined DeepSpeech and Google Cloud Speech and found that the systems were more likely to produce

errors around instances of habitual *be*, a grammatical feature of AAL. Following these results, Martin (2021) used corpus linguistics methods on four major spoken corpora used in training of ASR systems – Switchboard, Fisher, TIMIT, and LibriSpeech – and the Corpus of Regional African American Language (CORAAL). Martin found that habitual *be* is 1) far less frequent, 2) in fewer texts, and 3) surrounded by a less diverse set of word types and parts of speech in the four ASR corpora compared to CORAAL (2021). These findings showed that the spoken corpora used in training and evaluation of popular ASR systems are biased against AAL and "likely contribute to poorer ASR performance for Black speakers" (Martin 2021).

In contrast to this view, Koenecke et al. (2020) stated that racial disparities in ASR systems are primarily due to a performance gap in the acoustic models. This implies that the ASR systems are not adequately handling the phonological, phonetic, or prosodic features of AAL, as opposed to the grammatical or lexical features of AAL. Researchers do agree, however, that a likely cause of performance disparities is the lack of audio data from Black speakers in training ASR models (Koenecke et al. 2020; Martin and Tang 2020). Koenecke et al. tested this by using identical short phrases spoken by Black and White speakers as input to each of the five commercial ASR systems (2020). Their results showed similar performance disparities, suggesting that racial disparities are related to differences in pronunciation and prosody (Koenecke et al. 2020).

As speech recognition systems become more widespread, these racial disparities make it more difficult for African Americans to benefit from the technology. As Koenecke et al. note, these performance gaps have the potential to cause real harm to Black communities in cases where speech recognition systems are used "for example … by employers to automatically evaluate candidate interviews or by criminal justice agencies to automatically transcribe courtroom proceedings" (2020).

This thesis furthers research into racial disparities in ASR by examining the performance of two acoustic models on a forced alignment task of AAL speech data. Forced alignment is a task in speech recognition where, given an audio file and an orthographic transcription, the system creates a time-aligned transcript of the audio at the phoneme and word level. Notably, this task differs from ASR in that a transcript of the audio is required, meaning the system does not have to search for or predict which words are being said.

1.2 Experiments and Results

This work compares two acoustic models on a forced alignment task of AAL data. The first acoustic model comes from the Corpus of Regional African American Language (CORAAL). This acoustic model was trained on CORAAL version 2018.10.06, which included data from Black interviewees from Princeville, North Carolina and Washington, D.C. (Farrington and Kendall 2019). The second acoustic model used is a pretrained English acoustic model that is available through the Montreal Forced Aligner (MFA) and is trained on LibriSpeech, one of the four major spoken corpora examined in Martin (2021) that is trained on primarily white MUSE speech data. All other aspects of the speech recognition systems, including the pronunciation dictionary, the forced aligner system used, and the audio and transcriptions of AAL speech, were kept the same.

Each system was tasked with creating time-aligned transcripts of AAL data from Black speakers in Georgia. Two corpora were used, including CORAAL's Valdosta, Georgia interviews (Quartey et al. 2021) and data from the Roswell Voices component of the Linguistic Atlas Project (Kretzschmar 2016). The outputs from the two acoustic model systems were then compared to each other to analyze how each system performed on the AAL data. Several methods of analysis were used to target various features of AAL. Five distinctive consonantal features of AAL were chosen to be included in the analysis. The outputs of each system were analyzed to find

environments where these distinctive features could appear and determine to what extent the two systems showed the AAL variant of the feature in their output. The occurrences of the AAL variants in the system output were tabulated and compared across the two systems. Three vowel pairs that are of interest in AAL research were also chosen to be analyzed. After finding instances of the relevant vowels, the data from each system were compared to determine differences in vowel onset time, vowel duration, and vowel formant distribution. Vowel onset times were compared across systems by finding the difference between the two systems for each instance of a relevant vowel. The distributions of vowel durations were also compared across systems to determine if any vowels or speakers led to disagreement between the two systems. Pillai scores were used to determine the amount of separation in the distributions of two vowels in a vowel pair, and the Pillai scores were then compared across systems to determine if the two acoustic model systems' distributions differed.

The hypotheses for this study are in line with previous work on ASR performance disparities such as Koenecke et al. (2020), who suggested that a lack of variety in the ASR system training data was responsible for performance disparities. As the CORAAL acoustic model was trained on AAL speech data while the MFA acoustic model was not, the hypotheses for this work were as follows:

1) The system using the CORAAL acoustic model will produce a more accurate forced alignment of the AAL speech data than the control system that uses the MFA acoustic model
2) The CORAAL acoustic model system will be better able to distinguish vowels within each of the three vowel pairs than the control model

3) The CORAAL model will show more instances of phonetic transcriptions consistent with AAL phonological features.

The results of these analyses showed that the two acoustic model systems performed similarly in many of the environments analyzed. Vowel onset times were generally similar across systems, with consistent differences of less than 50 milliseconds being found in certain vowel contexts. The most disagreement between the two systems occurred in vowel duration, where statistically significant differences were found for every vowel pair. These differences indicate that the two systems differ in their ability to determine vowel durations, particularly with certain vowels. Differences were found in the Pillai scores of certain speakers in two vowel pair contexts, though most Pillai scores were comparable across systems.

1.3 Contributions

This thesis contributes to current research on racial disparities in ASR systems, focusing on the theory proposed by Koenecke et al. (2020) and others that acoustic models which lack training data from AAL speakers are a primary cause of disparities. This work examines whether an acoustic model that is trained exclusively on AAL data will perform better on a forced alignment task than a popular acoustic model trained on MUSE. Various common phonological features of AAL are analyzed in this work to determine which features and contexts may be more likely to lead to performance discrepancies. The results of this thesis show that, for a forced alignment task, both acoustic models perform similarly. These results indicate that when an exact transcript is available for AAL audio data, an AAL-specific acoustic model does not significantly improve performance. This research helps to illuminate the areas where speech recognition systems can effectively handle AAL data, namely forced alignment tasks. This work does not indicate that ASR systems in general perform well on AAL data, particularly with regard to performance metrics

such as Word Error Rate, but rather that if a complete transcript exists for AAL speech data, existing acoustic models can be used to effectively time-align those transcripts at the phone and word level.

1.4 Outline

The outline of the remainder of this thesis is as follows. Chapter 2 presents background information and previous research on AAL features, components of ASR systems, and the forced alignment process. The two corpora that are used are also described. Chapter 3 explains the preprocessing steps required to use MFA and how analyses of the outputs were conducted. Chapter 4 provides the results of the two systems with respect to the various analyses conducted. These include tabulations of the five AAL consonantal features, analysis of vowel onset times and vowel duration of the six vowels, and Pillai scores of the vowel pairs. The final chapter details the conclusions of this work by providing an overview of the results of analysis and what these results mean with respect to acoustic models and forced alignment of AAL, as well as directions for future work.

CHAPTER 2

BACKGROUND

2.1 African American Language (AAL)

The varieties of English spoken by African Americans have been referred to by different names throughout the past century, including Black English, Ebonics, African American English (AAE), African American Vernacular English (AAVE), and African American Language (AAL) (Kendall et al. 2018). AAL refers to all varieties of language use in African American communities and reflects differences within speakers' identities that intersect with ethnicity, race, and nationality (Lanehart and Malik 2015). AAL differs from other American English dialects in its phonological system, grammatical/morphosyntactic system, and its lexicon. While AAL does share features with varieties such as Mainstream U.S. English (MUSE), white Southern English, and Chicano English, AAL contains a unique combination of these and other features (Kendall et al. 2018).

This project uses several phonological features of AAL to analyze the performance of the forced aligner systems. These features can be generally split into consonantal features and vowel features. Five of the most well-studied phonological consonantal features of AAL were chosen to be included in the analysis of the speech recognition systems. The features were chosen based on their distinctiveness, meaning that they are salient features of AAL, and the regularity with which they appear in various varieties of AAL. These features and example instances are shown in Table

1 adapted from Lehr et al. (2014), Thomas and Bailey (2015), and ORAAL's "AAL Linguistic Patterns" web page[1] (Farrington and Kendall 2019).

Table 1: The five phonological features of AAL consonants chosen to be used in comparison of the two acoustic model systems, with a description and phonological rule mapping from MUSE to AAL realization and examples

Description	Phonological Rule	Examples
Reduction of word-final consonant clusters ending in [t] or [d]	C → ∅ / C _ #	hand → han' past → pas'
Devoicing of word-final voiced stops after a vowel, especially [d]	-cont, +voice] → [-cont, -voice] / V _ #	god → got
Dental fricative variation in [θ] as [t, f] and [ð] as [d,v]	[θ] → [t] or [θ] → [f] [ð] → [d] or [ð] → [v]	something → someting | tooth → toof other → udder | with → wiv
Deletion or vocalization of /l/ after a vowel	[l] → ∅ / V _ [l] → ə / V _	help → he'p fall → fauh
Deletion or vocalization of /r/ after a vowel or between two vowels	[r] → ∅ / V _ {#, V} [r] → ə / V _ {#, V}	father → fathuh here → heuh

With these features, it is important to note that 1) other varieties of English may exhibit similar features but do not exactly adhere to the AAL system and 2) regionality will affect which features an AAL speaker uses. For example, all varieties of American English show some level of consonant cluster reduction; Mainstream U.S. English (MUSE) and AAL both allow consonant cluster reduction when the following word begins with a consonant, as in *cold coffee → col' coffee* (Kendall et al. 2018). However, AAL also has a higher rate than MUSE of consonant cluster reduction where the following word begins with a vowel, as in *cold apple → col' apple* (Kendall et al. 2018). Similarly, English varieties such as Southern English and New York City English exhibit [r] dropping after a vowel or between vowels, but this feature has been diminishing in most

[1] https://oraal.uoregon.edu/AAL/Linguistic-Patterns

MUSE varieties since the mid-twentieth century while being maintained to some extent in AAL varieties (Sęn 1979). Additionally, as AAL is a purposefully general term that encompasses all varieties of language use in African American communities, the region that an AAL speaker is from may determine the features that the speaker does or does not use. In particular, Wolfram (1994) found that northern AAL speakers tended to delete or vocalize /r/ in fewer contexts than southern speakers. Hinton and Pollock (2000) built on this, finding that AAL speakers from Davenport, Iowa produced vocalic and postvocalic /r/ in all contexts compared to Memphis AAL speakers who showed consistent patterns of variation in their /r/ usage. As such, the features examined in this work are those which researchers would reasonably expect to find in southern AAL speakers.

Three vowel pairs were also chosen that represent features of AAL and Southern English varieties. This thesis uses IPA phonetic notation as well as the ARPABET phonetic transcription system to describe the vowels being examined. Table 2 shows the 2-letter ARPABET codes used for each vowel, the corresponding IPA symbol, and example words. ARPABET codes also typically include a number after the vowel to indicate no stress (0), primary stress (1), secondary stress (2), or tertiary and further stress (3). As unstressed vowels tend to be shorter and reduced (Lindblom 1963; Fourakis 1991), all vowel instances used in this analysis were primary stress vowels.

Table 2: A list of relevant ARPABET vowel codes with example words

ARPABET code	IPA symbol	Examples
AA	/ɑ/	bot, palm
AO	/ɔ/	bought, caught
IH	/ɪ/	bit, flint
EH	/ɛ/	bet, friend
IY	/i/	beat, eagle

Each vowel pair being examined shows some relevance to the African American Vowel Shift (AAVS) or general AAL in the existing literature (Thomas 2007; Kohn 2013). The AAVS is

a chain shift occurring in AAL that primarily affects the LOT, TRAP, DRESS, and KIT vowels, though Farrington et al. (2021) note that researchers should not expect all African Americans to be participating to the same extent in the AAVS due to how geographically dispersed African American communities are, demographic variation, and other social considerations. Whether the AAVS is more common in certain regions, age groups, and/or class backgrounds is an ongoing research question (Farrington et al. 2021).

The first vowel pair examined in this work targets the low back merger where /ɑ/ and /ɔ/ are no longer distinguished in words like *lot* and *thought*. AAL has traditionally kept a distinction between these two vowels (Labov et al. 2006; Bernstein 1993; Thomas 1989). There is also evidence of /ɑ/ fronting in AAL as part of AAVS, which may help maintain this distinction (Thomas 2001). The second vowel pair examined targets the merging of /ɪ/ and /ɛ/, a feature of both the Southern Vowel Shift (SVS) and AAL where /ɛ/ is raised and fronted as in the *pin-pen* merger. There is evidence of AAL speakers in Charleston, South Carolina adopting this merger (Baranowski 2013). Lastly, the vowel pair /ɪ/ and /i/ is examined, as in *bit* and *beat.* While /ɪ/ is characterized as being more tensed and raised in both the AAVS and the SVS, SVS also shows simultaneous laxing and falling of /i/ which is not present in AAVS (Holt 2016).

For the chosen vowel pairs, Pillai scores can be used to determine the amount of overlap between two vowels and are useful scores for ongoing vowel changes such as vowel mergers (Hall-Lew 2010; Hay et al. 2006). Pillai scores have been used to determine the effects of ongoing vowel shifts such as AAVS on AAL speakers' vowels (Renwick and Stanley 2017; Renwick and Olsen 2017, Newman et al. 2018) as well as the status of the ongoing vowel mergers analyzed in this work (Shi et al. 2019; Nycz and Hall-Lew 2013; Renwick and Stanley 2017).

In addition to Pillai scores, vowel onset and duration are known to be important features in automatic speech recognition. Prasanna et al. found that speech recognition performance on consonant-vowel units improved significantly when vowel onset points were used as an anchor point for feature extraction (2001). If vowel onsets are not accurately determined, performance of the overall speech recognition system will suffer as a result. Along with this, calculating a vowel's duration is also important to determine accurate formant values. Some variations in vowel duration occur in all varieties of English. For example, open vowels are typically longer in duration than closed vowels (Lehiste and Peterson 1961), tensed vowels are longer than lax vowels (Port and Rotunno 1979), and women typically produce longer vowels than men (Hillenbrand et al. 1995; Jacewicz et al. 2007). Outside of these, there are regional and dialectal differences in vowel duration. The speech of Southern U.S. speakers has been found to have longer vowels than northerners from New England, the Mid-Atlantic, and the West (Clopper et al. 2005) as well as those from central Ohio and southeastern Wisconsin (Jacewicz et al. 2007). With respect to AAL, Holt et al. found that Southern AAL speakers have significantly longer vowels than their white counterparts (2015). Holt et al. also reported that vowel duration did not differ as a function of age and that the tense-lax contrast was minimized for AAL speakers compared to MUSE speakers (2015). Increased vowel duration for AAL speakers was also found by Holt (2018), along with evidence that the AAL speakers in the study, from North Carolina, were not participating in the Southern Vowel Shift.

2.2 Automatic Speech Recognition

Automatic speech recognition (ASR) refers to the process by which a computer processes spoken speech into text. Most ASR systems use two components, an acoustic model and a language model, in conjunction with the overall system architecture. The acoustic model is trained on audio

data, taking the audio as input and outputting probabilities over phonetic units. Prior to deep learning, Gaussian mixture models (GMMs) and Hidden Markov Models (HMMs) were popular choices for acoustic models. GMMs output the most probable phone for a given time frame in the audio, but they do not use any phonetic context when determining the most probable phoneme. HMMs are temporal models where the architecture typically looks at three states of a phone – the beginning, middle, and end. Within this, each state is modeled by a GMM to determine the most likely phone. Recently, deep neural networks have begun to surpass GMM-based models (Yu et al. 2020). Current common approaches include a combination of deep learning and traditional methods, such as the DNN-HMM acoustic model that first helped promote deep learning applications in speech recognition tasks (Mohamed, Dahl, Hinton 2009).

The second component, the language model, is trained on text data to learn which word sequences are more likely to be produced in speech to aid in word prediction. The goal of the language model is to assign probabilities to words and phrases based on the training data. In short, the acoustic model contains the phonetic knowledge required for speech processing while the language model contains the knowledge of word, grammar, and syntactic structures of the language.

2.3 ASR Performance on AAL

Historically, AAL speakers have faced discrimination due to the stigmatization of AAL as ungrammatical speech. This is not an accurate representation of AAL, but AAL speakers continue to face discrimination in a multitude of ways. One area where this discrimination can be observed is in ASR. Koenecke et al. (2020) found that five major commercial ASR systems – Amazon, Apple, Google, IBM, and Microsoft – had substantially higher average word error rates (WER) for Black speakers than white speakers. Within each system, the WER for Black speakers was

almost double that of their white counterparts (Koenecke et al. 2020). Averaging the error rates across the five ASR systems yielded an aggregate WER of 0.25 for the Black speakers compared to 0.19 for the white speakers; even the ASR system with the best overall performance resulted in a WER of 0.27 for Black speakers compared to 0.15 for the white speakers (Koenecke et al. 2020). Additionally, Koenecke et al. found that WER was "strongly associated with AAVE dialect density," indicating that WER rises for Black speakers using these systems when they use more AAL features in their speech (2020). Le (2021) used the ASpIRE ASR model on the CallHome and CORAAL corpora and found the model performed significantly worse on Black speakers, in a way that interacted with their regionality; WERs were most impacted for AAL speakers who also used more features of Southern English. Additionally, Le examined a subset of words from the corpora that showed common AAL phonological features, finding that WER increased in the words containing AAL features, but only for Black speakers (2021). Tatman and Kasten (2017) examined the accuracy of Bing Speech and YouTube's automatic captions across race and found that both systems had lowest error rates for white speakers and higher error rates for African American and mixed race speakers. For speakers from the Pacific Northwest, Wassink et al. (2022) found that AAL speakers had higher normalized error frequency than Caucasian American speakers when using a custom-built ASR system, CLOx, used for sociolinguistic analysis. Usage of grammatical features of AAL can also lead to performance disparities. Martin and Tang (2020) evaluated DeepSpeech and Google Cloud Speech and found that instances of habitual *be*, a morpho-syntactic feature of AAL, and the surrounding words were more error prone than instances of non-habitual *be*.

2.4 Forced Alignment

Assuming an orthographic transcription exists for the speech data, the transcription can be aligned to the audio recording in a process called forced alignment, providing phone-level segmentation of the speech data. The inclusion of an initial transcript makes forced alignment different than other speech recognition tasks which would have to predict which word is being spoken based on the acoustic signal. Forced alignment is commonly used in cases where data exist but lack timeline information, such as movie transcripts, as well as being an important tool for phonological research. Manually transcribing and providing accurate time stamps for a movie script or linguistic interview is costly and prone to errors. Forced alignment solves these issues, but only if the results of the alignment are at least on par with human transcription.

To analyze the accuracy of a forced aligner, researchers typically compare the output to a human-annotated transcript (Goldman 2011; Coto-Solano et al. 2017; Gonzalez et al. 2018; MacKenzie and Turton 2020; Liu and Sóskuthy 2022). At the same time, there is research that supports the comparison of forced aligners without a human-annotated sample. MacKenzie and Turton, for example, found that aligner-placed and human-placed phoneme boundaries typically show only small displacements that would rarely have a significant effect on a researcher's measurements of interest (2020). Strunk et al. (2014) reported a mean inter-aligner difference of 85.5ms across eight samples from five languages. Goldman (2011) found that human vs. human agreement on phoneme boundary placement was roughly 80 percent at a 20ms threshold and 60 percent at 10ms for both English and French. Goldman also found that agreement between a machine alignment and human alignments was comparable to inter-human agreement (2011).

2.5 Montreal Forced Aligner

Various forced alignment systems exist, with one of the best-performing aligners being the Montreal Forced Aligner, an open-source system for speech-text alignment (McAuliffe et al. 2017;

Gonzalez et al. 2019). The Montreal Forced Aligner (MFA) is built off of the Kaldi ASR toolkit and uses a standard GMM/HMM architecture (McAuliffe et al. 2017). MFA consists of four primary training stages that use a combination of monophone models, which are context-independent, and triphone models, which consider the preceding and following phonetic context for each phone during training (McAuliffe et al. 2017). First, monophone GMMs are trained iteratively and used to generate a basic alignment. Next, triphone GMMs are trained to account for the surrounding phonetic context and generate new alignments. The triphone GMM alignments are then used to learn acoustic feature transforms for each speaker in the audio. Specifically, MFA uses Mel-frequency cepstral coefficients (MFCCs) as acoustic features. MFA calculates 13 MFCCs for 25 ms intervals using a 10 ms frame shift as well as delta and delta-delta features from surrounding MFCC frames, totaling 39 features per frame. In the third stage, cepstral mean and variance normalization (CMVN) is applied to the features on a per-speaker basis. Lastly, feature space Maximum Likelihood Linear Regression (fMLLR) is used to estimate feature transforms for each speaker (McAuliffe et al. 2017). The output of MFA includes phone-aligned TextGrid files and an acoustic model created from the data.

McAuliffe et al. (2017) first presented MFA and evaluated it against two widely-used aligners, FAVE and Prosodylab-Aligner. They report that MFA's architecture and retraining ability improved accuracy compared to FAVE and Prosodylab-Aligner (McAuliffe et al. 2017). Gonzalez et al. (2019) also found that MFA produced higher quality alignments than FAVE and MAUS, another common aligner. In some cases, the results of MFA alignment were not significantly different from human alignment (Gonzalez et al. 2019).

Included within MFA are two pretrained English acoustic models, English (US) ARPA acoustic model v2.0.0 and English MFA acoustic model v2.0.0. The first model is labeled as a

"General American English" dialect, which will be referred to here as Mainstream U.S. English (MUSE), and is trained on the LibriSpeech English corpus which contains roughly 1,000 hours of audiobooks from the LibriVox project (Panayotov et al. 2015). The second model contains training data from multiple English dialects but is predominantly trained on a combination of MUSE, British English, and Nigerian English. For the purposes of this project, the first model was chosen to serve as a control system since it was trained only on MUSE and uses the ARPABET phonetic transcription code, matching the dictionary used for the project.

2.6 CORAAL Acoustic Model

2.6.1 General Acoustic Model Creation

Acoustic models are created by taking audio recordings of speech and transcripts of the audio and compiling them into statistical representations of the phonemes. Feature extraction techniques are typically used to create the statistical representations (Bhatt et al. 2020). There are also various classification methods that can be used, which can be categorized as utilizing acoustic-phonetic knowledge, pattern recognition, artificial intelligence, or a combination of techniques (Bhatt et al. 2020; Saon and Chein 2012). Today, various toolkits exist to create a custom acoustic model or use a pretrained model, each with their own techniques and parameters (Lamere et al. 2003; McAuliffe et al. 2017; MacLean 2018). Acoustic models can thus be expected to differ in ability based on these factors. For acoustic models built using MFA such as CORAAL's acoustic model, the acoustic model is created as part of the forced alignment process outlined in section 2.5.

2.6.2 Creation of CORAAL's Acoustic Model

The Corpus of Regional African American Language (CORAAL), the first public corpus of AAL data, includes recorded speech from over 150 sociolinguistic interviews with speakers of

regional AAL varieties from multiple locations in the U.S. (Farrington and Kendall 2019). In 2018, CORAAL developed a new acoustic model using MFA and official CORAAL transcripts from CORAAL version 2018.10.06. This version included data from Princeville, North Carolina and two sets of interviews from Washington, D.C. (Farrington and Kendall 2019).

The first Washington, D.C. component (DCA) contains data from 68 speakers across 74 recordings collected between March 1968 and August 1969 as part of Ralph Fasold's work on AAL. The interviews reflect a Labovian sociolinguistic interview, including topics such as games, school, and favorite movies (Kendall et al. 2018). For CORAAL's purposes, speakers were selected from Fasold's interviews to best represent four age groups and three social class groups (Kendall et al. 2018).

The second D.C. component (DCB) comes from interviews conducted for CORAAL between July 2015 and December 2017, consisting of 48 primary speakers across 63 audio files (Kendall et al. 2018). Speakers for this component were collected through a friend network to fill a 4x3 demographic matrix, the same as for the DCA component. The interviews were sociolinguistic-styled interviews on topics including life in the D.C. area as well as the interviewee's neighborhood, schooling, and work history (Kendall et al. 2018).

The third component of CORAAL version 2018.10.06 that was used to create the new acoustic model includes data from Princeville, North Carolina. This component consists of 16 primary speakers across 32 audio files that were collected by Ryan Rowe, Walt Wolfram, and colleagues as part of the North Carolina Language and Life Project (Rowe et al. 2018). Princeville is the oldest town in the U.S. incorporated by African Americans, and many members of the community can trace their family lineage back to the original town founders. As of the 2000 census, 97% of the Princeville population identified as African American. The speakers in these

interviews were recorded between August 2003 and June 2004, with speakers selected to fill a 2x3 demographic matrix. The interviews were sociolinguistic styled interviews on topics including life in Princeville, schooling, and 1999 Hurricane Floyd, which flooded most of the town (Rowe et al. 2018). Based on the analysis of Le (2021), which included the Princeville component of CORAAL, the Princeville data represent speech that exhibits strong AAL and Southern features.

MFA alignment was first completed using version 2018.10.06, resulting in a new acoustic model created from these data as part of MFA's output. In 2019, CORAAL was re-aligned using the new acoustic model (Farrington and Kendall 2019).

CHAPTER 3

DATA AND METHODOLOGY

This chapter details the process of using the two acoustic model systems to complete forced alignment of two AAL speech corpora. The two corpora are detailed in sections 3.1 and 3.2, followed by acquisition of the corpora and the preprocessing steps that were undertaken before forced alignment in 3.3 and 3.4. Section 3.5 describes the process of using the Montreal Forced Aligner. Lastly, 3.6 includes a description of the various analyses performed on the output of MFA from both systems.

3.1 CORAAL Valdosta Corpus

As part of CORAAL's ongoing corpus-building endeavor, interviews that took place in Valdosta, Georgia were added to CORAAL version 2021.07 (Quartey et al. 2021). This component consists of 12 primary speakers across 14 audio files collected specifically for CORAAL from 2017 to 2019 in Valdosta, the county seat of Lowndes County in southern Georgia. The population of Valdosta is 56,000, with approximately 53% of the population identifying as African American as of the 2019 US Census estimate (Quartey et al. 2021). The 12 speakers were interviewed to fill a 3x2 demographic matrix as shown below in Table 3. The interviews are sociolinguistic styled interviews focusing on topics such as life in Valdosta, the interviewees' personal histories, and high school sports (Quartey et al. 2021). The Valdosta interviews were not added to CORAAL until 2021, and as such were not used in creating the 2018 acoustic model or re-aligning of CORAAL using the acoustic model in 2019.

Table 3: A 3x2 demographic matrix of the Valdosta interviewees, showing the number of male and female speakers across three age groups

	Female	Male
Age group 2 (20 to 29)	2	1
Age group 3 (30 to 50)	2	3
Age group 4 (51 and over)	2	2

3.2 Roswell Voices corpus

The Roswell Voices corpus is a component of the Linguistic Atlas Project (Kretzschmar et al. 2004; Kretzschmar et al. 2006; Andres and Votta 2009; Kretzschmar 2016). The data include 70 speakers from field work starting in 2002 in Roswell, Georgia, a city of roughly 95,000 people in northern Fulton County and a close suburb of Atlanta. African American and non-African American speakers are included in the data. Sonja Lanehart was the interviewer for the African American speakers while she was a faculty member at the University of Georgia. More recently, some of the interviews were manually transcribed and put into the TextGrid format that was used as input for the forced aligners in this work (Stanley et al. 2022). Only the African American speaker were used in this project. Note that not all of the African American speakers in the dataset have been transcribed, so only those with transcribed interviews were used here. This subset of the data is shown in Table 4 and includes 8 speakers across 14 recordings, with four males and four females.

Table 4: The name, gender, and number of interviews for the eight Roswell interviewees

Speaker	Gender	Number of Interviews
ROSWELL_INF006	Female	2
ROSWELL_INF021	Female	2
ROSWELL_INF023	Female	2
ROSWELL_INF024	Female	2
ROSWELL_INF009	Male	1

ROSWELL_INF011	Male	2
ROSWELL_INF017	Male	2
ROSWELL_INF042	Male	1

The data include guided conversational interviews that centered around social life in Roswell, fixed format elicitation where interviewees read words from cards, and 24 direct lexical questions (Kretzschmar 2016). The audio recordings are structured so that most speakers have two recordings, one of the guided conversational interview and a second of the fixed format elicitation and direct lexical questions. Two of the male speakers have only one recording each that contains the guided conversation interview.

3.3 Data Acquisition

The CORAAL Valdosta files were downloaded from the University of Oregon's Corpus of Regional African American Language website. The downloaded files included the audio recordings as .wav files and time-aligned TextGrid files at the utterance level.

The Roswell files were downloaded from the UGA GACRC. The downloaded files included the audio recordings as .wav files, TextGrid files time-aligned at the sentence level, and various metadata files for each speaker.

3.4 Preprocessing

A preprocessing script was created using Python which all TextGrid files went through. This initial preprocessing served several purposes. First, any rows of data that were labeled as being spoken by the interviewer rather than the interviewee were removed. This ensured that no speech data from non-AAL-speaking interviewers would influence the results of forced alignment. Since the TextGrid files included a start and end time for each utterance, the audio files were also preprocessed using the Python script so that audio from anyone other than the interviewee was removed and would not go through forced alignment. Next, flags for redacted information, portions labeled as unintelligible, and non-speech sounds such as laughter were removed. Most

punctuation was removed, with the exception of apostrophes, which were left in, and dashes and hyphens, which were normalized to each be one hyphen.

While there are countless other text preprocessing techniques that could have been utilized, this method was sufficient for the forced alignment task. More granular approaches are typically used when a human-annotated transcript exists for the data so as to not artificially inflate the word error rate (WER). However, since neither dataset contains ground-truth annotations, meaning annotations created by a human, at the word or phone level, this was not a concern for this project.

In terms of audio preprocessing, apart from removing non-interviewee portions of audio, no changes were made to the files during preprocessing. The CORAAL data did not require any changes to audio formatting. The Roswell audio files were encoded with a bit depth of 24, while MFA (via Kaldi) requires a bit depth of 16. However, from the MFA documentation[2], since higher bit depths are becoming more common for recording, MFA automatically converts higher bit depths to Kaldi's required 16. As such, no changes were made to the bit depth during preprocessing.

3.5 Using the Montreal Forced Aligner

All alignments were completed using MFA version 2.0 in a Linux environment. MFA requires four inputs: the audio files in a .wav format; transcripts of the audio as either TextGrid, .lab, or .txt files; a pronunciation dictionary; and the acoustic model. The pronunciation dictionary used was English (US) ARPA dictionary v2.0.0, which uses the ARPABET phone set and contains slightly under 200,000 words (Gorman et al. 2011). Two acoustic models were used. The first was the pretrained English acoustic model from the MFA website and described in 2.3, English (US) ARPA acoustic model v2.0.0. The second acoustic model was CORAAL's acoustic model

[2] https://montreal-forced-aligner.readthedocs.io/en/latest/user_guide/corpus_structure.html?highlight=bit%20depth#bit-depth

described in 2.5 and trained on AAL speech data from CORAAL's Washington, D.C. and Princeville, North Carolina data. For both the Valdosta and Roswell datasets, the audio and transcriptions were run through MFA twice, once using the pretrained MFA acoustic model and once using CORAAL's acoustic model. The same pronunciation dictionary was used for every alignment, and all materials used the ARPABET phone set. The total run time for the four forced alignments was approximately one hour.

Each forced alignment results in several output files. Most important for this project are the resulting word and phone-aligned TextGrid files for each speaker. The output also includes a list of words that were out of vocabulary (OOV), meaning they were not listed in the pronunciation dictionary but appeared in the original dataset.

3.6 Analyses

After all forced alignment tasks were completed, the resulting word and phone-aligned TextGrids were analyzed using Python and Praat to find differences between the performance of the two acoustic models, separated by dataset. All analyses compared the output of the two systems for each speaker. It is important to note that for both the Roswell and Valdosta datasets, no human-annotated phone or word-level transcriptions currently exist. This means that there is no gold standard to compare against the results of forced alignment. As there is previous research supporting the comparison of forced aligners without a human-annotated sample, the decision was made to analyze the output of the two systems with respect to vowel onset and duration time and compare the systems to examine performance differences in place of a comparison to a human annotator. This method places the focus of the comparison on finding environments where the two systems perform differently, rather than determining which system's output is more accurate.

First, each speaker's outputs by the two systems were compared using Python to find instances where the systems' phone-level representations of a word differed. To do this, each word of a transcript was added to a list of arrays, along with each system's resulting phonetic transcription of that word. Within each array, the two transcriptions could then be compared and, if they did not match, output to a CSV file. Next, Python's regular expression package was used to search for environments where an AAL feature could appear. The resulting words and phonetic transcriptions were saved, and the phonetic transcriptions were then analyzed to determine if they exhibited the AAL phonetic realization of a feature. Examples of the AAL features being examined, example words, and potential phonetic realizations using ARPABET are shown in Table 5. These instances were tabulated to determine if the two systems differed in the number of occurrences found for each consonantal feature.

Table 5: Descriptions of the five AAL features, example words, and AAL realizations

Feature	Word	AAL Realization
Reduction of word-final consonant clusters ending in [t] or [d]	Hand	H AE N
Devoicing of word-final voiced stops after a vowel, especially [d]	God	G AA T
Dental fricative variation in [θ] as [t, f] and [ð] as [d,v]	Tooth Other	T UW F AH D ER
Deletion or vocalization of /l/ or /r/ after a vowel	Tall Sister	T AO S IH S T AX

Provided in the output of MFA are the onset times and durations for each phoneme. The onset times and durations of the six vowels were separated and compared across systems. To analyze the vowel duration output, box plots were created to show the spread of vowel durations

for each speaker, separated by system. In place of a gold standard, these results were compared to existing literature on vowel duration (see Chapter 2.1).

For the onset time comparisons, the onset times for the relevant vowels were taken from the output, separated by speaker. This resulted in two lists, one for each system's output. Each item in the list contained the predicted vowel, onset time, the word the vowel appeared in, and the speaker. The two lists were then compared to each other to ensure that each instance of a vowel had a matching instance in the other system. To determine if an instance in one system had a matching instance in the other system, the two instances needed to examine the same vowel, in the same word, spoken by the same speaker. Additionally, the onset times reported by each system had to be within 1 second of each other to ensure the two instances were indeed of the same vowel instance. This method is not ideal as it doesn't account for potential onset time differences of more than one second. However, this was a necessary step to ensure that the same instances of a vowel were being compared across the systems.

This difference between two vowel onset times was calculated by subtracting the MFA control system's onset time from the CORAAL acoustic model system's onset time. If the resulting difference was positive, then the MFA control system's onset was earlier than the CORAAL system. Likewise, if the difference was negative, then the CORAAL system showed an earlier onset time than the MFA control system. These differences were used to create box plots to show the distribution of onset times by system for each vowel. Additionally, the differences were analyzed for each speaker to determine if the two forced alignment systems disagreed more for particular speakers.

For the three vowel pairs examined, a Praat script was used to find the formant values for each instance of the vowels. This was done using Praat's To Formant command with the following

parameters: time step: 0, maximum number of formants: 5, maximum hertz: 5000 for male speakers and 5500 for female speakers, window length: 0.025, and dynamic range: 30. Formant measurements were taken at the vowel midpoint and saved to output files for each speaker. These formant values were then compared across systems to determine if the two systems' formant values differed for certain vowels or certain speakers.

Pillai scores were also used in analyzing the vowel pairs. Pillai scores are used to describe the separability, or amount of overlap, of two distributions. Scores range from 0 to 1, with a score of 0 indicating total overlap and a score of 1 indicating completely distinct categories with no overlap. This score was chosen as the vowel pairs being analyzed all show some degree of overlap in both intraspeaker and interspeaker contexts. Shi et al. (2019) also used Pillai scores to examine the *pin-pen* merger in Southern speakers from the Digital Archive of Southern Speech (DASS), supporting its use here to examine vowel mergers, including *pin-pen*, in Southern AAL speakers. An R script was created to calculate each speaker's Pillai scores for a given vowel pair. This script was run twice for each dataset to calculate Pillai scores from each system's output. The following phoneme was also included as an independent variable when calculating the Pillai scores to account for the phonetic context in which a vowel appeared. The scores were calculated for each of the three vowel pairs, with the IH-EH vowel pair being split into prenasal and non-prenasal occurrences. Prenasal occurrences were also removed from the IH-IY data before calculating Pillai scores. Pillai scores for male and female speakers in each dataset were also calculated by averaging all speakers' individual Pillai scores.

After calculating Pillai scores, the decision was made to exclude the Roswell speakers' second interviews from the Pillai score analysis. Due to the format of the Roswell interviews, each speaker's second interview was much shorter than the first, ranging from 7 to 10 minutes. These

second interviews only included reading from a word list and direct lexical questions. After examining the Pillai scores from the second Roswell interviews, it was determined that some of the speakers may have altered their speech style in their second interview. That is to say, due to the format of this second interview, it is possible that the interviewees were more aware of their speech patterns while reading from the word list and were adopting a more MUSE speech style. This type of style shifting has been studied and found in African American communities and classic sociolinguistic interviews, being linked to the topic of speech, audience, and speaker identity (Garner and Rubin 1986; Craig et al. 2014; Grieser 2019). As a result, the Pillai scores for these second interviews appeared to show differences in the speakers' vowel space that were not present in their first interview. The Pillai scores for the Roswell speakers' second interviews were thus discarded. Note that this only affected the second interviews of Roswell speakers; a few Valdosta speakers were interviewed twice, but the second interviews there were much longer and were not affected by outliers.

CHAPTER 4

RESULTS

4.1 Consonantal Features

Five AAL consonantal features were chosen to examine differences in alignment from the two acoustic models. These features were 1) reduction of word-final consonant clusters ending in [t] or [d], 2) devoicing of word-final voiced stops after a vowel, 3) variation of dental fricatives [θ] and [ð] as [t, f] or [d, v], 4) deletion or vocalization of /l/ after a vowel, and 5) deletion or vocalization of /r/ after a vowel. For /l/ deletion or vocalization, /r/ deletion or vocalization, dental fricative variation, and word-final devoicing, no instances of these features were found in the phone output of either system, for either dataset. This is not to say that no speaker utilized these features in their speech, but rather that neither ASR system had any instances of using the phonological aspects of these features in their phonemic outputs.

To better understand this lack of AAL realizations, the pronunciation dictionary used during forced alignment, English (US) ARPA dictionary v2.0.0, was searched using regular expressions to determine if AAL phonetic realizations were included as potential pronunciations. The results of these searches showed that the pronunciation dictionary does not include realizations that would occur as a result of utilizing the four AAL features listed above. For example, the only phonetic realization for the word *thing* in the pronunciation dictionary is TH IH1 NG, disallowing potential variations such as T IH1 NG or F IH1 NG. Due to this, the pronunciation dictionary used during forced alignment effectively suppressed AAL realizations of these features as it did not allow either system to report the AAL variants in their output, even if speakers were utilizing these

features. No AAL-specific pronunciation dictionaries currently exist, and even well-known pronunciation dictionaries such as the CMU Pronouncing Dictionary do not include most AAL phonetic features. This is a clear direction for future work, as performance of the two acoustic models cannot be appropriately compared due to the limitations of the pronunciation dictionary.

In contrast to the other four AAL features, the English (US) ARPA dictionary v2.0.0 does include phonetic realizations of words where word-final consonant cluster reduction is allowed in AAL, such as *aroun'* for *around.* Instances of word-final consonant cluster reduction were found in both systems and in both datasets. In these instances, the phonemic outputs did not include the final [t] or [d], indicating that the consonant cluster was reduced and the [t] or [d] was not pronounced. Comparing between the CORAAL model and the control model, there is no significant difference in the number of instances found for this feature. While the two systems do differ slightly, overall performance with regard to this feature is the same. Additionally, the set of words which showed this feature were almost exactly the same in both the Roswell and Valdosta data; only one word, *playground*, was unique to the Roswell data. Tables 6 and 7 show the tabulation of these results.

Table 6: Tabulation of occurrences of consonant cluster reduction in the Roswell data.

Roswell Dataset

Word	Total number of occurrences	CORAAL system occurrences with deletion	MFA control system occurrences with deletion	Proportion of CORAAL system occurrences with deletion	Proportion of control system occurrences with deletion
around	67	53	61	0.79	0.91
last	28	16	17	0.57	0.61
most	79	39	28	0.49	0.35
thousand	11	8	8	0.73	0.73

second	6	3	5	0.5	0.83
playground	1	1	1	1.0	1.0
Total	192	120	120	.63	.63

Table 7: Tabulation of occurrences of consonant cluster reduction in the Valdosta data.

Valdosta Dataset

Word	Total number of occurrences	CORAAL system occurrences with deletion	MFA control system occurrences with deletion	Proportion of CORAAL system occurrences with deletion	Proportion of control system occurrences with deletion
around	76	59	65	0.78	0.86
last	61	34	31	0.56	0.51
most	103	35	30	0.34	0.29
second	28	19	18	0.68	0.64
thousand	18	13	17	0.72	0.94
Total	286	160	161	.56	.56

4.2 Vowel Onset Times

The phone-level MFA output includes onset times for each phoneme as well as the duration of the phoneme. A human-annotated transcript of the data at the phone level would provide a gold standard to compare each system against. Lacking this, the decision was made to find the difference between the two systems' onset times for each instance of a relevant vowel. This was done by subtracting the onset time of the MFA control model output from the onset time of the CORAAL model output, consistent with previous analyses of forced aligners (Goldman 2011; MacKenzie and Turton 2020; Gonzalez et al. 2020). If the resulting difference was a positive number, then the MFA control model had an earlier onset time than the CORAAL model. Likewise, if the resulting difference was negative, then the CORAAL model had an earlier onset time for that vowel. While this method cannot determine which system's output is closer to the true onset time (which would still be subject to human-annotator variation), it can still provide

useful data by showing general trends in the systems and specific areas that lead to more disagreement between the systems.

4.2.1 Individual Speaker Onset Time Differences

Average differences in reported onset times from the two systems were also calculated for each individual speaker. From this, it can be determined if the two systems disagreed more for certain speakers, as well as the number of speakers who had onset time differences across systems for each vowel. Table 8 shows where the two systems differed by an average of more than 10 milliseconds for all speakers. This shows that the systems disagree more for some speakers than others. The speaker labeled VLD_se0_ag3_m_01_1, a male Valdosta speaker, has onset time differences of at least 10 milliseconds for every vowel except IY. This speaker is also the only speaker from either dataset to have differences of over 40 milliseconds. The two largest differences for this speaker are for IH, both in a prenasal and non-prenasal context. Interestingly, these two large differences are opposite of each other; for prenasal IH, the CORAAL system reported later onset times than the control system, while for non-prenasal IH, the CORAAL system reported earlier onset times.

Table 8 also shows which vowels are most likely to differ in onset times and which system tends to show earlier onset times. Of the 28 speaker interviews, AO onset times differed across systems by an average of at least 10 milliseconds for 12 speaker interviews. In these 12 instances, the MFA acoustic model system reported earlier onsets in the majority of speakers. Onset times for prenasal IH also differed considerably for 11 of the 28 speaker interviews. The CORAAL acoustic model system typically reported earlier onsets for prenasal IH. Prenasal EH also showed differing onset times between the systems in 11 speaker interviews, and similar to prenasal IH, the CORAAL acoustic model system reported earlier onset times for the majority of the 11 instances.

Table 8: Average onset time differences for each speaker between the two systems, calculated using CORAAL_onset – MFA_onset. The number of check marks indicates the average difference, with each check mark equal to 10 milliseconds. A negative sign before the check mark(s) indicates that the result is negative.

Speaker Interview	AA	AO	IH	EH	IH prenasal	EH prenasal	IY
VLD_se0_ag2_f_01_1		✓	✓		- ✓	✓	
VLD_se0_ag2_f_02_1	- ✓				- ✓	- ✓	✓
VLD_se0_ag3_f_01_1	- ✓						
VLD_se0_ag3_f_01_2				- ✓✓		- ✓	
VLD_se0_ag3_f_02_1	✓	✓			- ✓✓		
VLD_se0_ag4_f_01_1						- ✓	✓✓
VLD_se0_ag4_f_02_1	- ✓	✓			- ✓✓✓	✓✓	
VLD_se0_ag2_m_01_1		- ✓			- ✓		✓
VLD_se0_ag3_m_01_1	✓	✓✓	- ✓✓✓✓	✓✓	✓✓✓✓	- ✓✓✓	
VLD_se0_ag3_m_01_2		✓✓✓	- ✓			- ✓✓✓	
VLD_se0_ag3_m_02_1		✓✓	- ✓✓			- ✓	
VLD_se0_ag3_m_03_1		✓		✓		- ✓✓	
VLD_se0_ag4_m_01_1		✓					
VLD_se0_ag4_m_02_1	- ✓			- ✓	✓✓		
ROSWELL_INF006_1							
ROSWELL_INF021_1		- ✓			✓		
ROSWELL_INF023_1					- ✓		
ROSWELL_INF024_1					- ✓✓		
ROSWELL_INF009_1						- ✓✓✓	
ROSWELL_INF011_1							
ROSWELL_INF017_1		✓✓	✓		- ✓✓✓		✓
ROSWELL_INF042_1	- ✓✓	✓		- ✓✓		- ✓✓✓	
Total: 28	7	12	5	5	11	11	4

4.2.2 Average Onset Time Differences in Valdosta and Roswell

After finding all onset time differences using the method described above, pairwise t-tests were conducted for each vowel to determine if any of the vowels differed significantly across systems. Female and male speakers were separated for each t-test.

For the female Valdosta speakers, prenasal IH had a statistically significant difference between the two systems with an average difference of -0.012 seconds ($p < .01$). This difference is negative, meaning that the CORAAL acoustic model system had earlier onsets than the MFA acoustic model system by an average of 12 milliseconds. No other vowels showed significant difference in onset times between the systems for the Valdosta female speakers.

For the male Valdosta speakers, both IY and prenasal EH show significant differences in onset times between the two systems. For prenasal EH, the average difference was -0.008 seconds, ($p < .05$), meaning that the CORAAL system had earlier onset times than the control system by an average of 8 milliseconds. The results for IY show the opposite trend; the CORAAL system reported later onset times for IY by a mean of 5 milliseconds ($p < .05$). While statistically significant, the average differences for both IY and prenasal EH here are quite small at less than 10 milliseconds.

For Roswell females, like in Valdosta, prenasal IH showed a statistically significant difference in onset times between the two systems. The mean difference for prenasal IH was -0.009 seconds ($p < .05$), indicating that CORAAL acoustic model system reported earlier onsets by an average of 9 milliseconds.

Lastly, comparing the onset times for the male Roswell speakers, prenasal IH showed a significant difference in onset times. CORAAL's acoustic model system reported earlier onsets for prenasal IH by an average of 12 milliseconds ($p < .05$).

4.2.3 Analysis of All Onset Time Differences

From the pairwise t-tests and table of onset time differences for each speaker, some patterns begin to emerge. Average onset times for prenasal IH were significantly different for Roswell males and females as well as Valdosta females. Prenasal EH showed significantly different average

onset times for the Valdosta male speakers. The CORAAL acoustic model system appears to find earlier onset times for prenasal IH and EH compared to the MFA control model. This result is somewhat unexpected as prenasal vowels are not typically expected to have an effect on the vowel's onset time. At the same time, it should be noted that even the largest differences in onset time between the two systems are under 50 milliseconds, indicating that even when the two systems disagree, their outputs are still quite similar. In comparison, Goldman (2011) and Raymond et al. (2002) found human annotators to agree on phoneme boundary at a rate of roughly 80 percent at a 20ms threshold. In other words, two human annotators can be expected to place phoneme boundaries within 20ms of each other roughly 80 percent of the time. The results of comparing the two acoustic model systems appear to be in line with this.

4.3 Vowel Duration Analysis

The reported duration of vowels can also be compared between the two forced alignment systems. Means were calculated to determine the average duration of each vowel for each speaker, and box plots were created to compare the two forced alignment systems. Pairwise t-tests were conducted to determine if differences in vowel duration were statistically significant. Second interviews with Roswell speakers were removed as the interviews were much shorter and highly affected by outliers. From the remaining data, the area where the two systems showed the most disagreement was in duration of prenasal IH. Significant differences were found for both male and female speakers in both Valdosta and Roswell. Average differences for these groups ranged from -6 to -16 milliseconds, indicating that the CORAAL acoustic model system reported shorter durations than the control system. Figure 1 shows the $\log_{10}$ distribution of prenasal IH durations for female Valdosta speakers as determined by both systems. For almost every speaker, the distribution is higher in the control system. This pattern also appears in duration for AA. Average

duration differences for AA ranged from -5 to -11 milliseconds, again indicating that the control system reported longer duration times than the CORAAL acoustic model system.

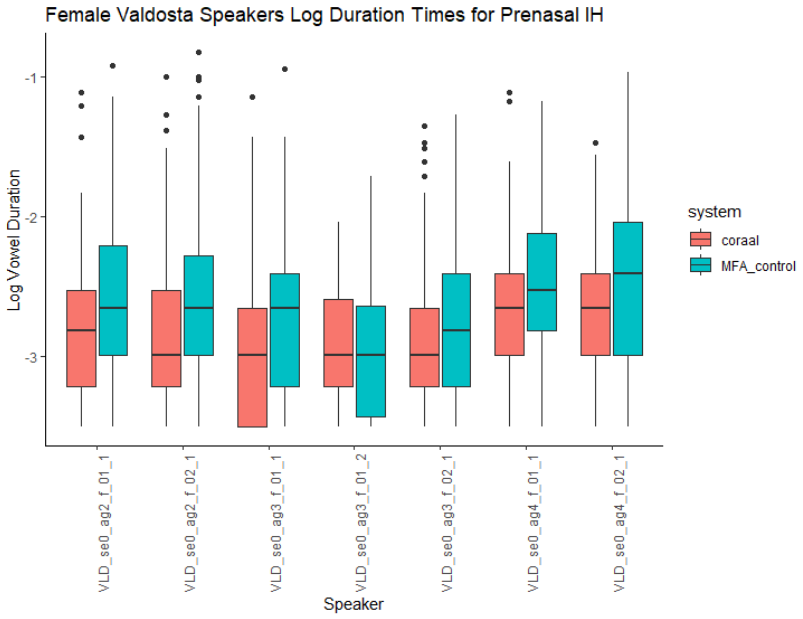

Figure 1: Box plots showing the log distribution of prenasal IH durations as found by each system for female Valdosta speakers

The vowel durations for IY show the opposite trend. Here, average durations were longer in the CORAAL acoustic model system with a range of 3 to 8 milliseconds. While these differences are smaller, they are statistically significant and in the opposite direction to prenasal IH and AA. Figure 2 shows the $\log_{10}$ distribution of IY durations for male Valdosta speakers, who had the highest average difference between systems of 8 milliseconds.

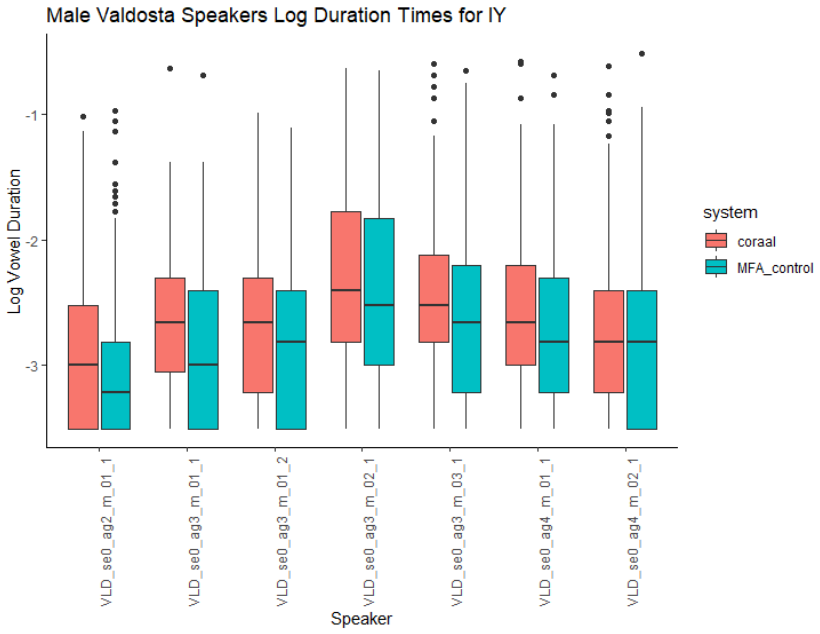

Figure 2: Box plots showing the log distribution of IY durations as found by each system for male Valdosta speakers

Overall, the most consistent average differences in vowel duration were found for AA, prenasal IH, and IY. Table 9 shows the average difference between the two systems for male and female speakers from each location as calculated by pairwise t-tests as well as the level of significance. Empty cells in Table 9 are areas where no significant differences were found. These results show that for AA and prenasal IH, the CORAAL acoustic model system reported shorter average duration times than the control model in every group. Non-prenasal IH also showed significantly shorter duration times from the CORAAL model in three of the four groups. As stated before, CORAAL reported significantly longer duration times for IY. For AO, significant differences were found for both groups of male speakers, though the differences are quite small.

Table 9: Average vowel duration differences between the two systems in milliseconds. A negative difference indicates that the CORAAL acoustic model system had shorter average duration than the control model. A positive difference indicates that the CORAAL system had longer average durations.
*Significance codes: * $p < .05$; ** $p < .01$; *** $p < .001$*

	AA	AO	Non-Prenasal EH	Non-Prenasal IH	Prenasal IH	Prenasal EH	IY
Valdosta females	-7			-6	-16 ***		4
Valdosta males	-5	6	4		-6		8 ***
Roswell females	-11 ***		-2	-8	-16 ***		5
Roswell males	-6	2	3	-4	-11 ***	2	3

There are a few possible explanations for these results. It is possible that the control acoustic model system is not accurately representing prenasal vowels, and the vowels are marked as having a longer duration due to the system's inability to distinguish the end of the prenasal vowel and beginning of the nasal consonant. On the other hand, it is also possible that the CORAAL acoustic model system is inaccurately representing prenasal vowels by prematurely assuming the prenasal vowel has ended and the nasal consonant has begun. This could also help to explain the results in 4.2 where the CORAAL acoustic model system reported earlier onsets for prenasal vowels.

While this analysis cannot determine which system's output is more correct, it is known that tense vowels such as IY typically have longer durations than lax vowels (Port and Rotunno 1979), which may be indicative of the CORAAL acoustic model finding more accurate duration times for this vowel. The CORAAL acoustic model system also reported shorter duration times for the lax vowel IH, which we can reasonably expect to have a shorter duration. At the same time, however, the CORAAL acoustic model system reported shorter durations than the control model for the tense vowel AA. It is also important to note that the largest difference between the two systems' outputs for average vowel duration was a difference of only 16 milliseconds. An analysis

of each speaker's vowel durations found that both systems reported very similar mean durations for each individual with system differences of less than 5 milliseconds.

4.4 Individual Pillai Scores

Pillai scores were used to examine each of the three vowel pairs targeting ongoing vowel mergers and features of AAVS – 1) AA and AO for the *cot-caught* merger, 2) IH and EH for the *pin-pen* merger, and 3) IH and IY, where a more tensed and raised IH and similarly tense and raised IY would indicate participation in the AAVS. Here, the Pillai scores are calculated using the formant measurements of each vowel, which were collected at the midpoint of each vowel. Depending on how the two systems calculated measures like the time points for vowels, these formant values may be different, leading to different Pillai scores. It is important that a forced alignment system correctly represents the speaker's vowel space so that researchers can draw accurate conclusions about which linguistic features a speaker is or is not exhibiting. This is especially relevant for vowels that are undergoing changes, such as the ones analyzed in this thesis which relate to the AAVS.

As Kennedy (2006) notes, one limitation of the Pillai score is that it lacks a meaningful measure of significance along with the distance measurement. Pillai scores are designed so that the score is more likely to be significant if the two distributions are clearly distinct (Kennedy 2006). Hall-Lew (2010) further describes that the range of Pillai scores across a speaker sample can be used to represent the relative extent of merger between any two speakers. However, the MANOVA will only identify speakers with clearly distinct vowels and cannot provide statistical discrimination between speakers with near-mergers and complete mergers (Hall-Lew 2010). In this work, the two acoustic model systems are considered to have a noticeable difference in Pillai scores for a vowel pair if the two systems' scores differ by at least .10.

Pillai scores for each individual speaker were calculated for both systems, for each vowel pair. The two systems were largely in agreement, with most speakers' Pillai scores differing by 0.05 or less between the two acoustic model systems (see Table 10). However, there were a few speakers whose scores varied by more than 0.10 across systems. Of these speakers, each had only one or two vowel pairs where Pillai scores differed considerably across systems. Additionally, these differences were limited to only the AA-AO vowel distinction and the non-prenasal IH-EH distinction. No significant differences were found between the two systems for IH-IY or prenasal IH-EH vowel pairs.

Table 10.a: Pillai scores for each speaker in the Valdosta dataset, for each of the vowel pairs and contexts examined.

Speaker	CORAAL AA AO	Control AA AO	CORAAL IH EH (prenasal)	Control IH EH (prenasal)	CORAALIH EH (non-prenasal)	Control IH EH (non-prenasal)	CORAAL IH IY	Control IH IY
VLD_se0_ag2_f_01_1	0.456	0.446	0.158	0.155	0.32	0.322	0.144	0.229
VLD_se0_ag2_f_02_1	0.156	0.21	0.0245	0.0242	0.254	0.32	0.223	0.171
VLD_se0_ag2_m_01_1	0.307	0.331	0.115	0.0684	0.251	0.261	0.099	0.0943
VLD_se0_ag3_f_01_1	0.381	0.444	0.0128	0.0314	0.284	0.251	0.151	0.175
VLD_se0_ag3_f_01_2	0.369	0.416	0.0258	0.0317	0.381	0.444	0.238	0.291
VLD_se0_ag3_f_02_1	0.291	0.361	0.0288	0.036	0.387	0.397	0.145	0.125
VLD_se0_ag3_m_01_1	0.266	0.519	0.0749	0.036	0.138	0.0622	0.101	0.15
VLD_se0_ag3_m_01_2	0.25	0.371	0.0514	0.0234	0.344	0.263	0.0758	0.0977
VLD_se0_ag3_m_02_1	0.474	0.599	0.0724	0.123	0.4	0.304	0.15	0.159
VLD_se0_ag3_m_03_1	0.382	0.388	0.0538	0.00598	0.418	0.416	0.0943	0.0725
VLD_se0_ag4_f_01_1	0.528	0.522	0.0796	0.0773	0.273	0.297	0.128	0.158
VLD_se0_ag4_f_02_1	0.335	0.333	0.00923	0.0243	0.313	0.335	0.159	0.111
VLD_se0_ag4_m_01_1	0.251	0.323	0.0671	0.0581	0.268	0.222	0.205	0.195
VLD_se0_ag4_m_02_1	0.369	0.352	0.145	0.0664	0.175	0.182	0.057	0.0535

Table 10.b: Pillai scores for each speaker in the Roswell dataset, for each of the vowel pairs and contexts examined.

Speaker	CORAAL AA AO	Control AA AO	CORAAL IH EH (prenasal)	Control IH EH (prenasal)	CORAALIH EH (non-prenasal)	Control IH EH (non-prenasal)	CORAAL IH IY	Control IH IY
ROSWELL_INF006_1	0.44	0.34	0.165	0.115	0.37	0.23	0.362	0.444
ROSWELL_INF021_1	0.276	0.261	0.00666	0.0204	0.34	0.24	0.21	0.273

ROSWELL_INF023_1	0.445	0.441	0.0709	0.108	0.314	0.256	0.389	0.403
ROSWELL_INF024_1	0.472	0.437	0.0662	0.109	0.35	0.418	0.346	0.296
ROSWELL_INF009_1	0.139	0.227	0.0219	0.0487	0.269	0.247	0.372	0.397
ROSWELL_INF011_1	0.284	0.37	0.0432	0.0386	0.352	0.27	0.174	0.257
ROSWELL_INF017_1	0.109	0.187	0.0343	0.0184	0.243	0.167	0.209	0.22
ROSWELL_INF042_1	0.09	0.09	0.08	0.08	0.37	0.26	0.28	0.21

4.4.1 Valdosta Speakers

From the Valdosta dataset, one male speaker showed Pillai score differences of 0.10 across systems for the non-prenasal IH-EH distinction, and two male speakers had Pillai score differences of more than 0.10 for the AA-AO distinction. Note that Valdosta speaker VLD_se0_ag3_m_01 participated in two interviews, with the interview number appearing at the end of the speaker label. Looking at non-prenasal IH-EH, the CORAAL acoustic model system resulted in a Pillai score of 0.40, while the MFA acoustic model system resulted in a score of 0.30, indicating that the CORAAL system showed slightly less overlap between non-prenasal IH and EH than the control system for this speaker. Two speakers also showed differences in Pillai scores for the AA-AO distinction. One of these speakers participated in two interviews and showed differences in both interviews. For the three instances of differences in AA-AO Pillai scores, the CORAAL acoustic model system resulted in a lower Pillai score than the MFA control system (see Table 11). This indicates that the MFA acoustic model system showed less overlap between these speakers' AA and AO vowels than the CORAAL acoustic model system. Figures 3 and 4 shows the output of both systems for one Valdosta speaker's AA and AO vowels. As the Pillai scores show, the CORAAL system shows greater overlap between the two vowels. The MFA acoustic model system displays a narrower distribution of AO vowels on the F2 axis compared to the CORAAL acoustic model system, where formant values are more scattered.

Table 11: Valdosta speakers whose Pillai scores differed by at least .10 across systems (shown in bold)

	AA-AO		Prenasal IH-EH		Non-prenasal IH-EH		IH-IY	
Valdosta Speaker	CORAAL	Control	CORAAL	Control	CORAAL	Control	CORAAL	Control
VLD_se0_ag3_m_01_1	**0.27**	**0.52**	0.08	0.04	0.14	0.06	0.10	0.15
VLD_se0_ag3_m_01_2	**0.25**	**0.37**	0.05	0.02	0.34	0.26	0.08	0.01
VLD_se0_ag3_m_02_1	**0.47**	**0.60**	0.07	0.12	**0.40**	**0.30**	0.15	0.16

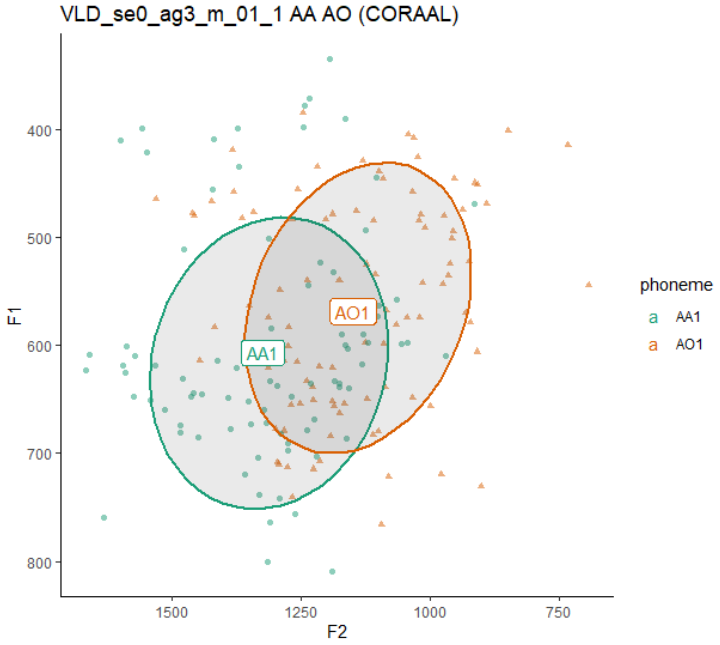

Figure 3: A male Valdosta speaker's AA and AO vowel instances in the CORAAL acoustic model system

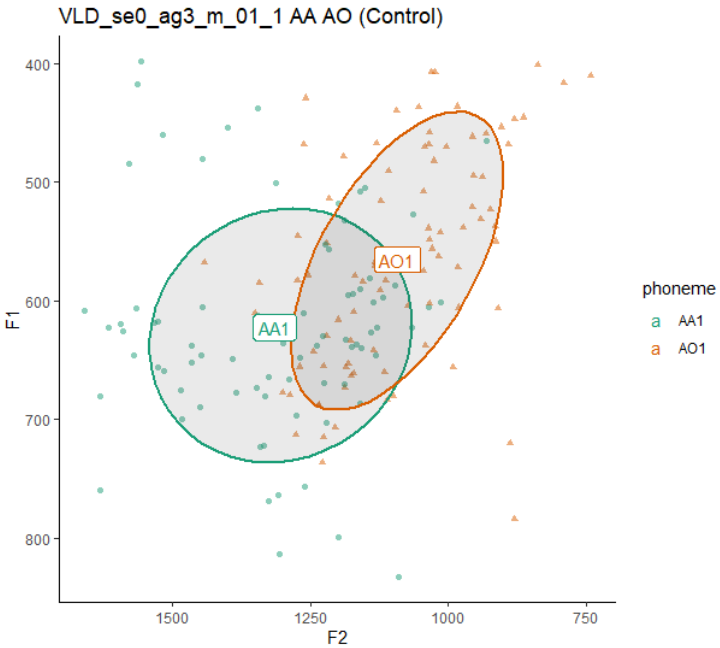

Figure 4: A male Valdosta speaker's AA and AO vowel instances in the control system

4.4.2 Roswell Speakers

From the Roswell data set, two female speakers and one male speaker had Pillai score differences of more than 0.10 across the two systems. All three speakers showed differences in their non-prenasal IH-EH vowels. One of the female speakers also had Pillai score differences in her AA-AO vowels. For all instances of non-prenasal IH-EH differences as well as the instance of AA-AO differences, the CORAAL acoustic model system reported a higher score than the MFA acoustic model system. The Pillai scores for all three speakers are shown in Table 12 with system differences of more than 0.10 in bold.

Table 12: Roswell speakers whose Pillai scores differed by at least .10 across systems (shown in bold)

	AA-AO		Prenasal IH-EH		Non-prenasal IH-EH		IH-IY	
Roswell Speaker	CORAAL	Control	CORAAL	Control	CORAAL	Control	CORAAL	Control
ROSWELL_INF006_1	**0.44**	**0.34**	0.17	0.12	**0.37**	**0.23**	0.36	0.44
ROSWELL_INF021_1	0.28	0.26	0.01	0.02	**0.34**	**0.24**	0.21	0.273
ROSWELL_INF042_1 (male)	0.09	0.09	0.08	0.08	**0.37**	**0.26**	0.28	0.21

For the non-prenasal IH-EH vowel distinction, all three speakers had higher degrees of vowel separation, or less overlap between the two vowel distributions, in the CORAAL acoustic model system. Figures 4 and 5 show the instances of female speaker ROSWELL_INF006_1's non-prenasal IH and EH vowels from the CORAAL and MFA control acoustic model systems, along with ellipses showing one standard deviation around the means for each vowel[3]. When looking at just the ellipses, there is little noticeable difference between the two acoustic model systems. By plotting each instance of the vowels, however, it becomes clearer that the CORAAL acoustic model system shows slightly less overlap between the two vowels. Both systems show general clusters for each vowel with a large overlap, but the MFA control acoustic model system appears

[3] This speaker shows a bimodal EH distribution, with clusters of higher and lower F1 tokens. The cluster closer to IH is predominantly pre-rhotic EH tokens. Both acoustic model systems display this similarly.

to have more instances of EH scattered amongst the IH vowel cluster compared to the CORAAL acoustic model system, consistent with the Pillai scores of 0.37 and 0.23 for the CORAAL and MFA acoustic model systems, respectively.

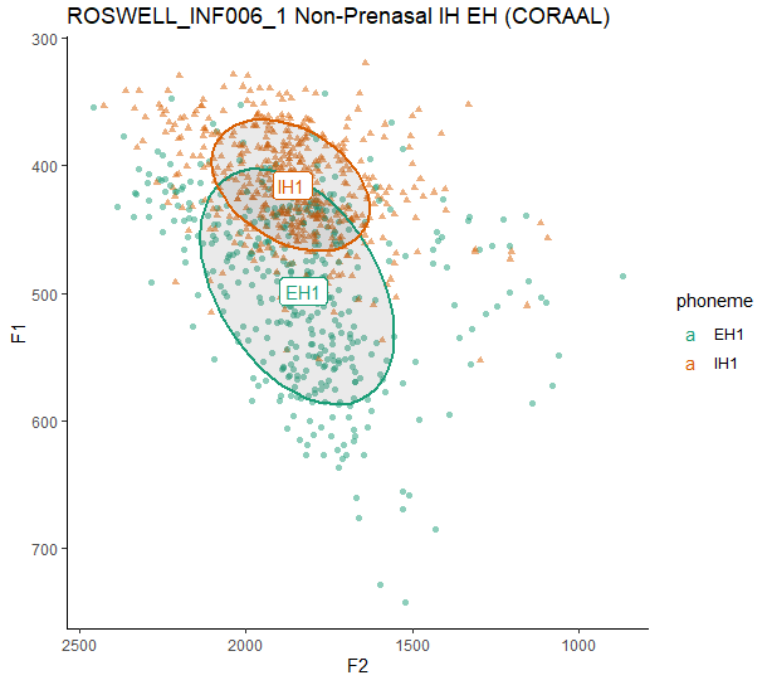

Figure 5: A female Roswell speaker's instances of non-prenasal IH and EH as determined by the CORAAL acoustic model system

Figure 6: A female Roswell speaker's instances of non-prenasal IH and EH as determined by the control acoustic model system

Female speaker ROSWELL_INF006_1 also showed a higher Pillai score from the CORAAL acoustic model system for AA and AO vowels. Here, the CORAAL acoustic model system resulted in a Pillai score of 0.44, while the MFA acoustic model system showed a score of 0.34. Figures 6 and 7 show the results of the two systems on this speaker's AA and AO vowels.

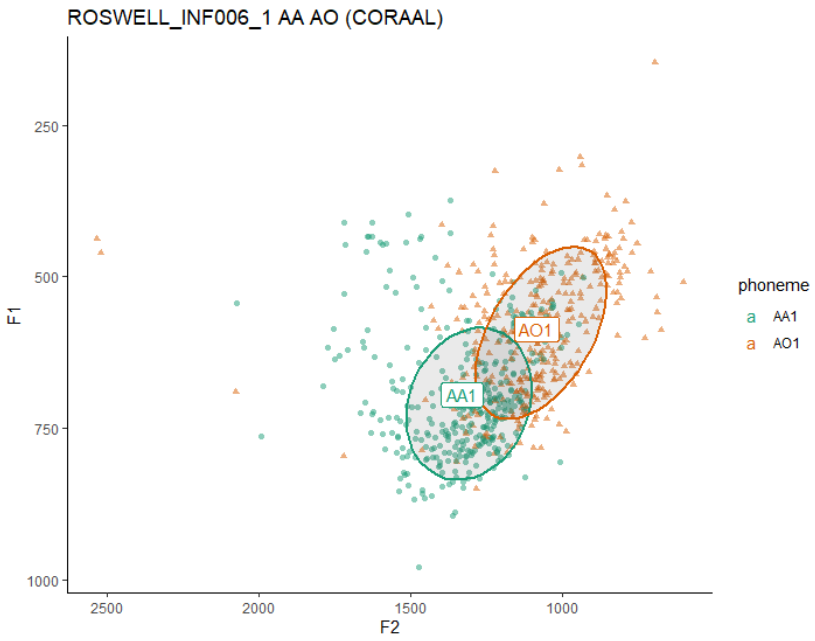

Figure 7: One Roswell female's AA and AO vowels as found by the CORAAL acoustic model system

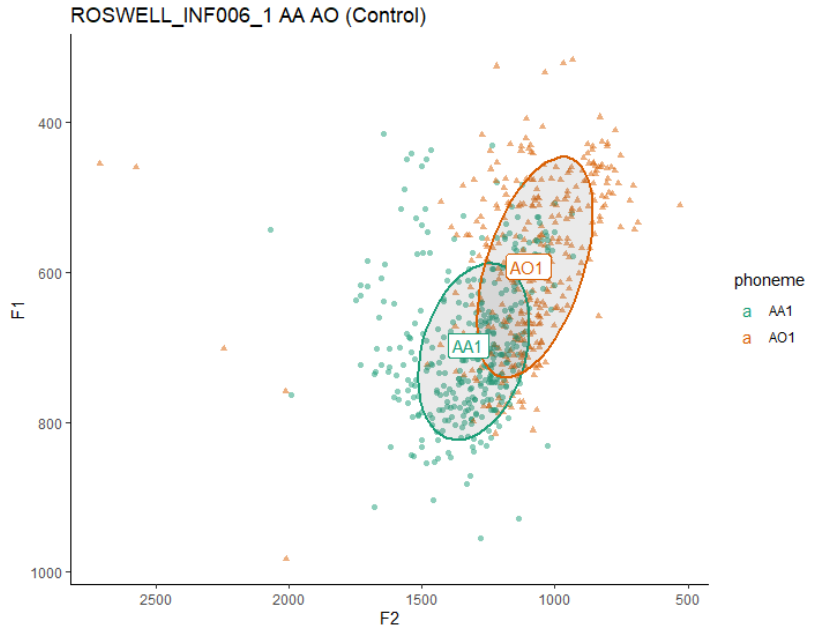

Figure 8: One Roswell female's AA and AO vowels as found by the MFA acoustic model system

In all four cases where the two acoustic model systems resulted in Pillai score differences of more than 0.10 for Roswell speakers, the CORAAL acoustic model system showed a higher degree of separation between the two vowel distributions. These results suggest that while the two systems perform very similarly in the majority of speakers, there are a few cases where the CORAAL acoustic model system finds more separation between vowels. In every instance of disagreement between the two systems with respect to non-prenasal IH and EH, regardless of speaker location, the CORAAL acoustic model system showed at least slightly more separated vowel distributions than the control model. However, the results for the AA-AO vowel pair are less clear. In the Valdosta data, the CORAAL acoustic model system always reported less separation of AA and AO compared to the control model. In the Roswell data, in the one instance of AA-AO difference, the CORAAL model reported more separation than the control model. This could be due to regional differences between Roswell and Valdosta.

4.5 Average Pillai Scores for Males and Females

After calculating each individual speaker's Pillai scores, average Pillai scores for male and female speakers in each dataset were calculated. The results, shown in Table 13, show that the two systems perform very similarly in the amount of separation between vowel pairs.

Table 13: Average Pillai scores for males and females for the vowel pairs AA-AO, IH-EH, and IH-IY, split by dataset and alignment system. Numbers in parentheses indicate the standard deviation from the mean.

Average Pillai Scores for Males and Females in Each Dataset for Each System				
	Valdosta		Roswell	
	CORAAL	MFA	CORAAL	MFA
Female AA-AO	0.36 *(0.12)*	0.39 *(0.08)*	0.41 *(0.09)*	0.37 *(0.09)*
Male AA-AO	0.33 *(0.10)*	0.41 *(0.11)*	0.16 *(0.09)*	0.22 *(0.12)*
Female prenasal IH-EH	0.05 *(0.05)*	0.05 *(0.03)*	0.08 *(0.07)*	0.09 *(0.05)*
Male prenasal IH-EH	0.08 *(0.05)*	0.05 *(0.04)*	0.04 *(0.03)*	0.05 *(0.03)*
Female non-prenasal IH-EH	0.32 *(0.05)*	0.34 *(0.11)*	0.34 *(0.02)*	0.29 *(0.09)*
Male non-prenasal IH-EH	0.28 *(0.06)*	0.24 *(0.11)*	0.31 *(0.06)*	0.24 *(0.05)*
Female IH-IY	0.17 *(0.04)*	0.18 *(0.05)*	0.33 *(0.08)*	0.35 *(0.08)*

Male IH-IY	0.11 *(0.06)*	0.12 *(0.05)*	0.26 *(0.09)*	0.27 *(0.09)*

The differences between the two systems are all within a range of 0.08, indicating broad agreement between the systems. Comparing between male and female speakers, Pillai scores for males are slightly lower for most of the vowel pairs, regardless of system. This is most noticeable in the Roswell AA-AO data. While female speakers in both locations do not show merged *cot/caught*, the Roswell male speakers appear more merged than the Valdosta male speakers, whose scores are in line with the female speakers. This suggests that the male Roswell speakers exhibit merged AA and AO vowels that the other speakers do not. Additionally, Valdosta speakers show lower Pillai scores for the IH-IY vowel pair compared to the Roswell speakers, indicating more overlap within the Valdosta speakers' vowel space. This result is consistent with vowel shifting from both the AAVS and SVS.

Figures 8 and 9 show plots of AA and AO for Roswell females and males with ellipses showing one standard deviation from the mean F1 and F2 values. The plots show that, for the AA-AO vowel distinction, the female Roswell speakers show a higher degree of separation between the two sounds. The female speakers AA vowels are slightly more fronted than the males, a feature of AAVS that can help maintain the AA-AO distinction. This is not to say that the Roswell females completely maintain the AA-AO distinction, as the Pillai scores do not support this. However, the females do show more separated AA and AO vowels than the male Roswell speakers. The output of the two acoustic model systems show strong similarities.

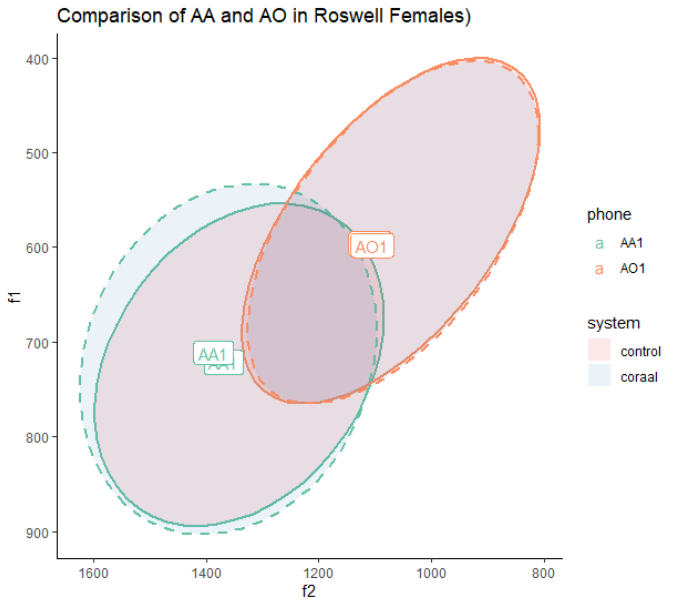

Figure 9: AA and AO plots for Roswell females with ellipses showing one standard deviation from the mean F1 and F2 values

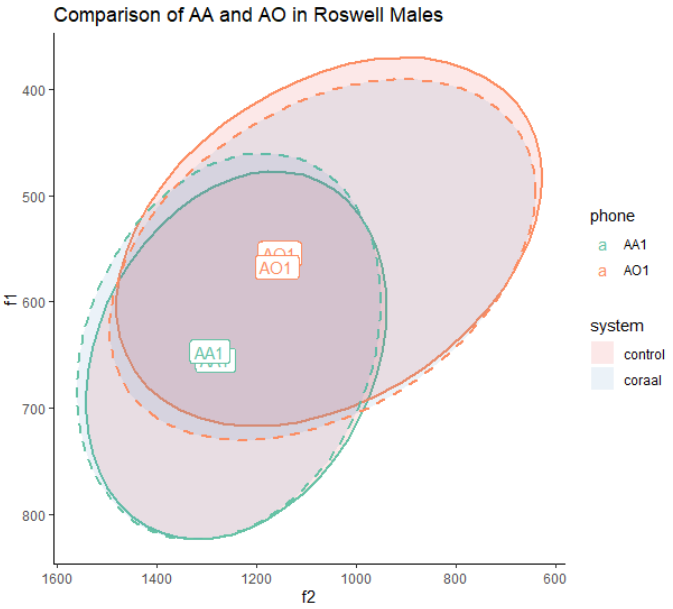

Figure 10: AA and AO plots for Roswell males

4.6 Summary

The output of forced alignment from both acoustic model systems was analyzed in regard to four areas: 1) consonantal AAL features, 2) vowel onset times, 3) vowel durations, and 4) merged status of vowels via Pillai scores. The analysis of consonantal features showed that for /l/ deletion or vocalization, /r/ deletion or vocalization, dental fricative variation, and word-final

devoicing, neither system had any instances of the AAL phonology for these features in their output. Analysis of the pronunciation dictionary used during forced alignment showed that the AAL phonetic realizations of these features are not included within the dictionary, meaning that even if speakers did utilize these features, the features could not be represented in the phonetic output of forced alignment. The pronunciation dictionary did, however, include AAL variants of word-final consonant cluster reduction, and both systems showed a similar number of instances of [t] and [d] deletion in a similar set of words. Vowel onset times were mainly in agreement, but were more likely to differ for certain speakers than others. The CORAAL acoustic model system tended to find earlier onset times for prenasal IH and EH. On the other hand, the CORAAL system found later onset times for IY in Valdosta males. These differences were statistically significant but small, with average differences of roughly 10 milliseconds. For vowel duration, the CORAAL acoustic model system consistently reported shorter durations for AA and prenasal IH and longer duration times for IY. These results were also statistically significant, with the greatest average difference between the systems being 16 milliseconds.

Analysis of individual speaker Pillai scores showed broad agreement across the systems. For two Valdosta males, the two systems differed in Pillai scores by more than .10 in four instances. Three of these were for the AA-AO vowel distinction, where the CORAAL acoustic model system reported lower Pillai scores, and the remaining instance was for the non-prenasal IH-EH distinction, where the CORAAL system reported a higher Pillai score. For Roswell speakers, Pillai scores differed by more than .10 across systems for three speakers, two females and one male. For the non-prenasal IH-EH distinction, the CORAAL acoustic model system again reported higher Pillai scores, and the CORAAL system also reported a higher Pillai score for one female's AA-AO vowel distinction. Lastly, Pillai scores were compared across the male and

female Valdosta and Roswell speakers. The Roswell males were unique in that they showed a more merged *cot-caught* than Roswell females or Valdosta speakers of either gender. Additionally, Pillai scores for the Valdosta speakers indicated more merged IH and IY vowels than the Roswell speakers. This is consistent with stronger participation in the AAVS and/or SVS by the Valdosta speakers.

Overall, differences were not found with respect to consonantal features of AAL. Vowel onset times were generally in agreement but were more likely to differ for certain speakers. Vowel durations were significantly different across systems, with the CORAAL acoustic model system finding shorter durations for AA and prenasal IH and longer duration times for IY, as compared to the control system. Pillai scores were broadly in agreement, though differences were found for several speakers in each dataset.

CHAPTER 5

CONCLUSION

The goal of this thesis was to analyze the performance of two acoustic models on a forced alignment task using AAL speech data. From the previous literature on racial disparities in ASR systems and how acoustic models may contribute to the issue, this thesis furthered work in the area by examining the following questions: 1) Will an acoustic model trained on AAL data perform better on an AAL forced alignment task compared to an existing widely-used, MUSE-trained acoustic model? 2) Does performance between the two acoustic models differ based on phonological environment, speaker, region, or gender? 3) To what extent will the two acoustic model systems include AAL realizations in their output? It was hypothesized that the CORAAL acoustic model system would produce a more accurate forced alignment of the AAL speech data, be better able to distinguish vowels within each of the three vowel pairs than the control model, and show more instances of phonetic transcriptions consistent with phonological features present in AAL.

5.1 Results and Trends

Five consonantal features of AAL were examined to determine if the two acoustic model systems differed in their treatment of common AAL features. Four of the five features did not appear in the output of either system for either dataset, as the pronunciation dictionary did not include these features as potential variants. Instances of word-final consonant cluster reduction consistent with AAL were found in both systems. These occurrences were similar across both

systems, with the feature being found in the same words and a similar number of times in each system.

Three vowel pairs that are often examined in AAL research were chosen to look for differences between the two systems. In regard to vowel onset times, the results of these analyses showed small differences in onset times for prenasal IH and EH in both Roswell and Valdosta. This result is somewhat surprising, as a nasal following a vowel would not typically be expected to influence the vowel's onset time. Additionally, the onset time analysis showed that certain speakers led to more disagreement between the systems. This could be due to differences in the extent to which speakers use AAL features.

The most consistent differences between the two systems occurred in vowel durations. Significant differences were found for AA and prenasal IH, where the CORAAL acoustic model system reported shorter durations, as well as IY, where the CORAAL system reported longer vowel durations. These differences were found for male and female speakers from both Valdosta and Roswell, indicating that this result is not due to a difference between regional AAL varieties used in Roswell and Valdosta. There are multiple ways to interpret this result, and further work should be done to determine which system is producing more accurate vowel durations for the AAL speakers. It is possible that an acoustic model trained on CORAAL data does a better job locating the offset of prenasal vowels, which could explain the differences in prenasal vowel onset time and duration. However, it is also important to remember that even when the differences between systems were statistically significant, the differences themselves were typically small.

Pillai scores were also used to determine the amount of overlap reported for a vowel pair for each system. For four Valdosta speakers and three Roswell speakers, the two systems differed in Pillai scores by more than .10. In the Valdosta data, three of these were for the AA-AO vowel

distinction, where the CORAAL acoustic model system reported lower Pillai scores, and the remaining instance was for the non-prenasal IH-EH distinction, where the CORAAL system reported a higher Pillai score. For the three Roswell speakers, the CORAAL acoustic model system again reported higher Pillai scores for non-prenasal IH and EH, and the CORAAL system also reported a higher Pillai score for one female's AA-AO vowel distinction. Average Pillai scores for male and female Valdosta and Roswell speakers were also compared. These showed that Roswell males exhibit a more merged *cot-caught* than female Roswell speakers or Valdosta speakers of either gender. Pillai scores for the Valdosta speakers were lower than the Roswell speakers for the IH-IY vowel distinction, consistent with both the AAVS and SVS.

5.2 Implications and Limitations

The results of this work show that the two acoustic models performed quite similarly but did show significant difference with relation to certain vowels on a forced alignment task of AAL speakers from Georgia. It is important to remember that forced alignment is a different task than ASR as a whole. Forced alignment assumes the existence of a full transcription of the audio data. In a task such as forced alignment, the speech recognition system does not need to predict which word is being spoken or search for potential words. Additionally, the transcriptions used as input for forced alignment typically have utterance-level time stamps in the TextGrid file, further simplifying the task of forced alignment. Previous work has shown that ASR systems do not perform well on AAL data but have not examined forced alignment of AAL speech. As the two acoustic models performed similarly on this task, it appears that the existence of a transcript of audio data is very helpful in processing AAL speech. The results also show that, in general, a dialect-specific pretrained acoustic model is not needed for forced alignment. This is helpful for linguistic researchers who study various dialects of a language as they generally can use a single

model trained on a general variety of the language. These results also support previous work such as MacKenzie and Turton (2020), who found that a model trained on MUSE was able to effectively perform forced alignment of British English varieties.

While in general the two systems performed similarly, the analysis of vowels relevant to ongoing AAL changes showed that the systems do disagree in significant ways in certain vowel contexts. It is possible that the data each acoustic model was trained on are responsible for these differences. This result would be in line with current theories proposed by Koenecke et al. (2020) and others. It could also be the case that the differences are caused by the way in which one or both systems determines the boundaries of phonemes. This would explain why differences were found in both the onset times and durations of certain vowels.

The largest limitation of this work is the lack of human-annotated data at the word and phone level. Without this, it is impossible to determine measures such as word error rate. The lack of gold-standard transcriptions also limited the methods that could be used to analyze vowel onset time and duration. Secondly, the lack of AAL realizations within the pronunciation dictionary used in this work limited the extent to which the systems were able to produce transcriptions consistent with AAL phonology. It is noted on the MFA website[4] that for pronunciation dictionaries, transcription accuracy and lexicon coverage typically cater to the prestige variety of a language, and this is certainly the case in this work. Only one of the five AAL features examined in this work, word-final consonant cluster reduction, appears as a potential phonetic variant.

5.3 Future Work

[4] https://mfa-models.readthedocs.io/en/latest/dictionary/English/English%20%28US%29%20ARPA%20dictionary%20v2_0_0.html#English%20(US)%20ARPA%20dictionary%20v2_0_0

The results of this thesis create various paths for future work. Perhaps most obvious is to tackle the lack of human-annotated data. This could be done by developing human annotations for the datasets used in this work or by utilizing different corpora that include human-annotated data. This would allow for better analysis of how the models perform with respect to a gold standard transcription. With the data used in this work, there is an open question of why some speakers appeared to have more discrepancies in the two systems' outputs. Koenecke et al. (2020) and Martin (2021) found that ASR system errors increased as more instances of AAL features were used. An analysis of the amount of AAL features used by the speakers in these datasets could help determine if this is the cause for performance discrepancies between speakers.

Another potential avenue for future work would be to adapt current pronunciation dictionaries to include pronunciations common in dialects outside of the prestige dialect. Previous work such as Shi et al. (2019), Bailey (2016), and Yuan and Liberman (2011) have shown the effectiveness of adding multiple pronunciations to a dictionary. Additions to the pronunciation dictionaries used by speech recognition systems would lead to greater dialectal coverage and higher accuracy for varieties that are not currently being represented. Including these pronunciation variants would also allow for a more accurate comparison of the two acoustic models, as the pronunciation dictionary would not be suppressing phonetic realizations that do not appear in the prestige variety of a language.

Lastly, CORAAL is one of the only large-scale corpora of AAL speech, and there are very few acoustic models or ASR systems trained on AAL data. Existing literature on ASR performance disparities points to a lack of non-MUSE training data (Tatman 2017, Koenecke et al. 2020, Martin 2021). Developing more datasets for non-MUSE speakers is an important next step. There is also an ongoing research question as to whether and to what extent ASR systems can perform multi-

dialectal speech recognition (Elfeky et al. 2016, Li et al. 2018, Chen et al. 2015). Future work could therefore work on developing and implementing more non-MUSE corpora in multi-dialectal speech recognition tasks.

REFERENCES

Andres, Claire & Rachel Votta. 2009. African American Vernacular English: Vowel Phonology in a Georgia Community. Publication of the American Dialect Society 94(1). 75–98. https://doi.org/10.1215/-94-1-75.

Bailey, George. 2016. Automatic Detection of Sociolinguistic Variation Using Forced Alignment. University of Pennsylvania Working Papers in Linguistics 22(2). http://repository.upenn.edu/pwpl/vol22/iss2/3.

Baranowski, M. (2013). "Ethnicity and Sound Change" by Maciej Baranowski. https://repository.upenn.edu/pwpl/vol19/iss2/2/

Bernstein, C. (1993). Measuring Social Causes of Phonological Variation in Texas on JSTOR. https://www.jstor.org/stable/455631

Bhatt, S., Jain, A., & Dev, A. (2020). Acoustic modeling in speech recognition: a systematic review. International Journal of Advanced Computer Science and Applications, 11(4).

Chen, M., Yang, Z., Liang, J., Li, Y., & Liu, W. (2015). Improving deep neural networks based multi-accent Mandarin speech recognition using i-vectors and accent-specific top layer. Interspeech 2015, 3620–3624. https://doi.org/10.21437/Interspeech.2015-718

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

Coto-Solano, R., & Solórzano, S. F. (2017). Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. CLEI Electron. J., 20(1), 2-1.

Craig, H. K., Kolenic, G. E., & Hensel, S. L. (2014). African American English–speaking students: A longitudinal examination of style shifting from kindergarten through second grade.

Elfeky, M., Bastani, M., Velez, X., Moreno, P., & Waters, A. (2016). Towards acoustic model unification across dialects. 2016 IEEE Spoken Language Technology Workshop (SLT), 624–628. https://doi.org/10.1109/SLT.2016.7846328

Farrington, C., & Kendall, T. (2019). CORAAL MFA-Aligned. http://lingtools.uoregon.edu/coraal/aligned/

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. The Journal of the Acoustical society of America, 90(4), 1816-1827.

Garner, T., & Rubin, D. L. (1986). Middle class Blacks' perceptions of dialect and style shifting: The case of southern attorneys. Journal of Language and Social Psychology, 5(1), 33-48.

Goldman, J. P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In Interspeech'11, 12th Annual Conference of the International Speech Communication Association.

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. Linguistics Vanguard, 6(1), 20190058. https://doi.org/10.1515/lingvan-2019-0058

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics, 39, 192.

Green, L. (2002). A descriptive study of African American English: Research in linguistics and education. International Journal of Qualitative Studies in Education, 15(6), 673–690. https://doi.org/10.1080/0951839022000014376

Grieser, J. A. (2019). Investigating topic-based style shifting in the classic sociolinguistic interview. American Speech, 94(1), 54-71.

Hall-Lew, L. (2010). Improved representation of variance in measures of vowel merger. 060002–060002. https://doi.org/10.1121/1.3460625

Hay, Jennifer, Paul Warren & Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. Journal of Phonetics 34(4). 458–484.

Hillenbrand, J., Getty, L. A., & Clark, M. J. (n.d.). Acoustic characteristics of American English vowels. 14.

Hinton, L. N., & Pollock, K. E. (2000). Regional variations in the phonological characteristics of African American Vernacular English. World Englishes, 19(1), 59-71.

Holt, Y. F. (2016). Sociophonetic analysis of vowel variation in African American English in the Southern United States. 060008. https://doi.org/10.1121/2.0000453

Holt, Y. F. (2018). Mechanisms of Vowel Variation in African American English. Journal of Speech, Language, and Hearing Research, 61(2), 197–209. https://doi.org/10.1044/2017_JSLHR-S-16-0375

Holt, Y. F., Jacewicz, E., & Fox, R. A. (2015). Variation in Vowel Duration Among Southern African American English Speakers. American Journal of Speech-Language Pathology, 24(3), 460–469. https://doi.org/10.1044/2015_AJSLP-14-0186

Jacewicz, E., Fox, R. A., & Salmons, J. (2007). Vowel Duration in Three American English Dialects. American Speech, 82(4), 367–385. https://doi.org/10.1215/00031283-2007-024

Kendall, T. (2018). ORAAL - AAL Linguistic Patterns. https://oraal.uoregon.edu/AAL/Linguistic-Patterns

Kendall, Tyler and Charlie Farrington. 2021. The Corpus of Regional African American Language. Version 2021.07. Eugene, OR: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal.

Kendall, T., Fasold, R., Farrington, C., McLarty, J., Arnson, S., & Josler, B. (2018). Corpus of Regional African American Language: Washington, D.C. A. http://lingtools.uoregon.edu/coraal/

Kendall, Tyler, Jason McLarty, and Brooke Josler. 2018. ORAAL: Online Resources for African American Language: AAL Facts. Eugene, OR: The Online Resources for African American Language Project. https://oraal.uoregon.edu/facts

Kendall, T., Quartey, M., Farrington, C., McLarty, J., Arnson, S., & Josler, B. (2018). Corpus of Regional African American Language: Washington, D.C. B. http://lingtools.uoregon.edu/coraal/

Kennedy, M. (2006). Variation in the pronunciation of English by New Zealand school children.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.

Kretzschmar, William A. 2016. Roswell Voices: Community Language in a Living Laboratory. In Karen P. Corrigan & Adam Mearns (eds.), Creating and Digitizing Language Corpora: Volume 3: Databases for Public Engagement, 159–175. London: Palgrave Macmillan UK.

Kretzschmar, William A., Jr., Becky Childs, Bridget Anderson, and Sonja Lanehart. 2004. Roswell Voices, Roswell: Roswell Folk and Heritage Bureau. [booklet and CD]

Kretzschmar, William A., Jr., Claire Andres, Rachel Votta, and Sasha Johnson. 2006. Roswell Voices, Phase 2, Roswell: Roswell Folk and Heritage Bureau. [booklet and CD]

Kohn, Mary Elizabeth. 2013. Adolescent ethnolinguistic stability and change: a longitudinal study. Chapel Hill, NC: The University of North Carolina at Chapel Hill Ph.D. Dissertation.

Kohn, M., & Farrington, C. (n.d.). A Tale of Two Cities: Community Density and African American English Vowels. 12.

Labov, W., Ash, S., & Boberg, C. (2008). The atlas of North American English: Phonetics, phonology and sound change. Walter de Gruyter.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., ... & Wolf, P. (2003, April). The CMU SPHINX-4 speech recognition system. In Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong (Vol. 1, pp. 2-5).

Lanehart, S. (2015). The Oxford Handbook of African American Language.

Le, L. (2021). Race and Regionality on the ASpIRE ASR Model (Doctoral dissertation, University of Georgia).

Lehiste, I., & Peterson, G. E. (1961). Transitions, Glides, and Diphthongs. The Journal of the Acoustical Society of America, 33(3), 268–277. https://doi.org/10.1121/1.1908638

Lehr, M., Gorman, K., & Shafran, I. (2014). Discriminative pronunciation modeling for dialectal speech recognition.

Li, B., Sainath, T. N., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., & Rao, K. (2018). Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4749–4753. https://doi.org/10.1109/ICASSP.2018.8461886

Lindblom, B. (1963). Spectrographic study of vowel reduction. The journal of the Acoustical society of America, 35(11), 1773-1781.

MacLean, K. (2018). Voxforge. Ken MacLean.[Online]. Available: http://www.voxforge.org/home.[Acedido em 2012].

Martin, J. L., & Tang, K. (2020). Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be." Interspeech 2020, 626–630. https://doi.org/10.21437/Interspeech.2020-2893

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

MacKenzie, Laurel & Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. Linguistics Vanguard. De Gruyter Mouton 6(s1). https://doi.org/10.1515/lingvan-2018-0061.

Mohamed, A., Dahl, G. E., & Hinton, G. (2012). Acoustic Modeling Using Deep Belief Networks. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 14–22. https://doi.org/10.1109/TASL.2011.2109382

Newman, M., Haddican, B., & Tan, Z. Z. G. (2018). Almost everyone in New York is raising PRICEs. University of Pennsylvania Working Papers in Linguistics, 24(2), 10.

Nycz, J., & Hall-Lew, L. (2013, December). Best practices in measuring vowel merger. In Proceedings of Meetings on Acoustics 166ASA (Vol. 20, No. 1, p. 060008). Acoustical Society of America.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7178964/

Port, R. F., & Rotunno, R. (1979). Relation between voice-onset time and vowel duration. The Journal of the Acoustical Society of America, 66(3), 654–662. https://doi.org/10.1121/1.383692

Prasanna, S., Gangashetty, S., & Yegnanarayana, B. (2001). Significance Of Vowel Onset Point For Speech Analysis. 81–88.

Quartey, M., Farrington, C., Kendall, T., Tacata, C., & McLean, J. (2021). Corpus of Regional African American Language: Valdosta, Georgia. http://lingtools.uoregon.edu/coraal/

Raymond, W. D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., & Hilts, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. In Seventh International Conference on Spoken Language Processing.

Renwick, M. E., & Olsen, R. M. (2017). Analyzing dialect variation in historical speech corpora. The Journal of the Acoustical Society of America, 142(1), 406-421.

Renwick, M. E., & Stanley, J. A. (2017, June). Static and dynamic approaches to vowel shifting in the Digital Archive of Southern Speech. In Proceedings of Meetings on Acoustics 173EAA (Vol. 30, No. 1, p. 060003). Acoustical Society of America.

Rowe, R., Wolfram, W., Kendall, T., Farrington, C., & Josler, B. (2018). Corpus of Regional African American Language: Princeville, North Carolina. http://lingtools.uoregon.edu/coraal/

Saon, G and J. T. Chien, “Large-vocabulary continuous speech recognition systems: A look at some recent advances,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 18–33, 2012, doi: 10.1109/MSP.2012.2197156

Sęn, Ann L. (1979). "English in the Big Apple: Historical Backgrounds of New York City Speech". The English Journal. 68 (8): 52–55. doi:10.2307/815156. JSTOR 815156

Shi, Y., Renwick, M. E., & Maier, F. (2019). Improved vowel labeling for prenasal merger using customized forced alignment. The Journal of the Acoustical Society of America, 146(4), 2957-2957.

Stanley, Joseph A., Jon Forrest, Lelia Glass and Margaret E. L. Renwick. Perspectives on Georgia vowels: From legacy to synchrony. Oral presentation at the 2022 Annual Meeting of the American Dialect Society. Washington, D.C. January 6 – 9, 2022.

Strunk, J., Schiel, F., & Seifart, F. (2014, May). Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS. In LREC (pp. 3940-3947).

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. (2001). An acoustic analysis of vowel variation in New World English. Publication of the American Dialect Society.

Thomas, E. R. (2007). Phonological and phonetic characteristics of African American vernacular English. Language and Linguistics Compass, 1(5), 450-475. https://doi.org/10.1111/j.1749-818X.2007.00029.x.

Thomas, E. R. (1989). Vowel changes in Columbus, Ohio. Journal of English linguistics, 22(2), 205-215.

Thomas, E. R., & Bailey, G. (2015). Segmental Phonology of African American English. The Oxford Handbook of African American Language, 403-419.

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. Speech Communication, 140, 50–70. https://doi.org/10.1016/j.specom.2022.03.009

Wolfram, W. (1968). ERIC - ED028423—A Study of the Non-Standard English of Negro and Puerto Rican Speakers in New York City. Volume I: Phonological and Grammatical Analysis., 1968. https://eric.ed.gov/?id=ED028423

Wolfram, W. (1969). ERIC - ED028431—A Sociolinguistic Description of Detroit Negro Speech. Urban Language Series, No. 5., 1969. https://eric.ed.gov/?id=ED028431

Wolfram, W. (1994). On the sociolinguistic significance of obscure dialect structures: The [NP i call NP i V-ing] construction in African-American Vernacular English. American Speech, 69(4), 339-360.

Yu, C., Kang, M., Chen, Y., Wu, J., & Zhao, X. (2020). Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview. IEEE Access, 8, 163829–163843. https://doi.org/10.1109/ACCESS.2020.3020421

Yuan, Jiahong & Mark Liberman. 2011. Automatic detection of “g-dropping” in American English using forced alignment. In 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, 490–493. IEEE. https://doi.org/10.1109/ASRU.2011.6163980.