An Investigation on Prenasal Merger in Southern American English Through Automatic Speech Recognition

by

Yuanming Shi

(Under the Direction of Frederick Maier)

Abstract

Commonly used in Automatic Speech Recognition programs, forced alignment is the process of generating phone-level timestamped transcriptions based on orthographic transcriptions. In this work, a modified pronunciation dictionary for forced alignment is used to identify the prenasal merger (a.k.a. pin-pen merger) in Southern US English. We hypothesize that the modification will increase the acoustic separation between the prenasal allophones of /ɪ/ and /ɛ/. This is borne out in our experiments performed on the Digital Archive of Southern Speech audio corpus. We compare vowel formant values before and after the modification, including the separation between vowel clusters as measured by Pillai scores and Euclidean distances between vowel centroids. K-means clustering on vowel formant values is used to show our modification yields better phonetic transcriptions. We also use Kullback-Leibler Divergence to show the trained acoustic models for the prenasal allophones of /ɪ/ and /ɛ/ differ the most in their final portions.

Index words: Phone Classification, Forced Alignment, Formant Extraction, Automatic Speech Recognition, Prenasal Merger, Gaussian Mixture Models, K-means Clustering

An Investigation on Prenasal Merger in Southern American English Through Automatic Speech Recognition

by

Yuanming Shi

B.A., Wuhan University, 2012
M.A., Western Michigan University, 2015
M.A., Rice University, 2017

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

2019

An Investigation on Prenasal Merger in Southern American English Through Automatic Speech Recognition

by

Yuanming Shi

Approved:	
Major Professor:	Frederick Maier
Committee:	Margaret Renwick Khaled Rasheed

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Methodology

Vowel mergers indicate sound changes where two or more contrastive vowels are replaced by a single vowel in a speaker's speech. The prenasal merger, which is a distinctive feature in Southern US English, is a merger in which speakers drop the distinction between IH (/ɪ/) and EH (/ɛ/) in words whose canonical pronunciations contain EH followed by a nasal consonant (including N, M, NG, which correspond to the IPA symbols of /n/, /m/, and /ŋ/ respectively).[1] For instance, when the merger is present, the word *pen* sounds like the word *pin*, but the word *pet* still sounds very different from the word *pit*. Early studies on this topic can be found in [1], [2], [3], among others. Although the prenasal merger is prevalent among speakers from the American South, the degree of the merger varies among these speakers and some of them may not have the merger at all.[2]

The general goal of this thesis is to use forced alignment to provide an analysis of pre-

[1] IH, EH, N, M, NG are ARPABET symbols. For a detailed description on the correspondence relation between APRABET and IPA symbols, see Table A.1 in Appendix A.

[2] [1] documents the evolution of the prenasal merger in Tennessee. Specifically, evidence shows that the merger has become predominant in the South only in this century". Moreover, later in this thesis, examples are given on the presence and absence among different Southern speakers. Discussions can be found in Section 2.1 and Figure 2.3.

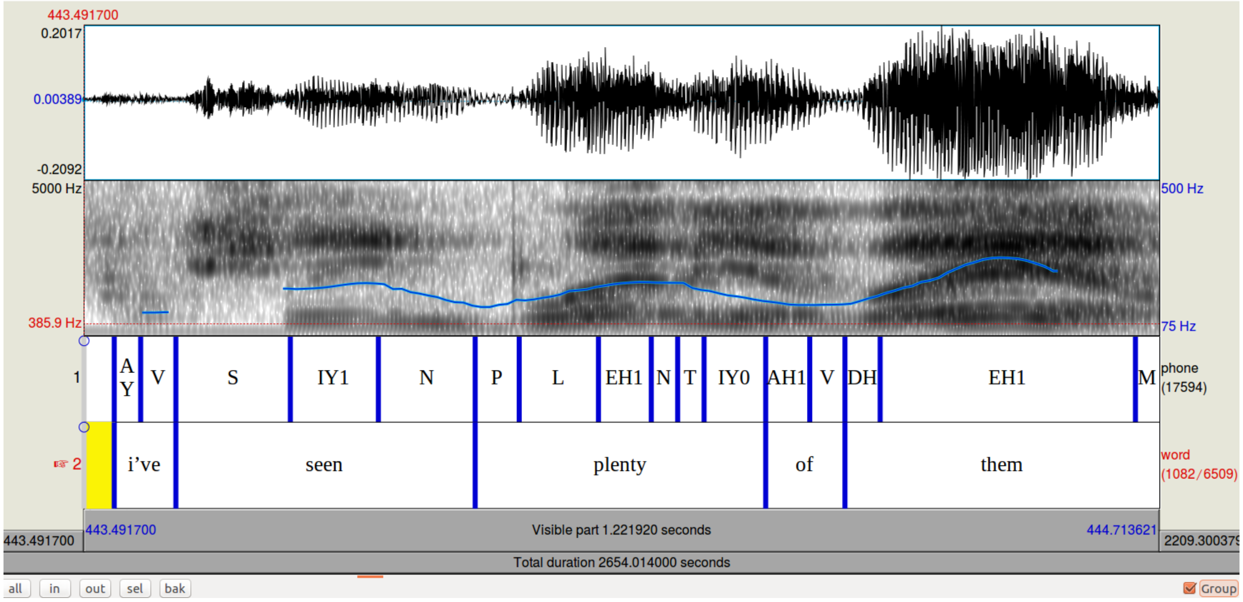
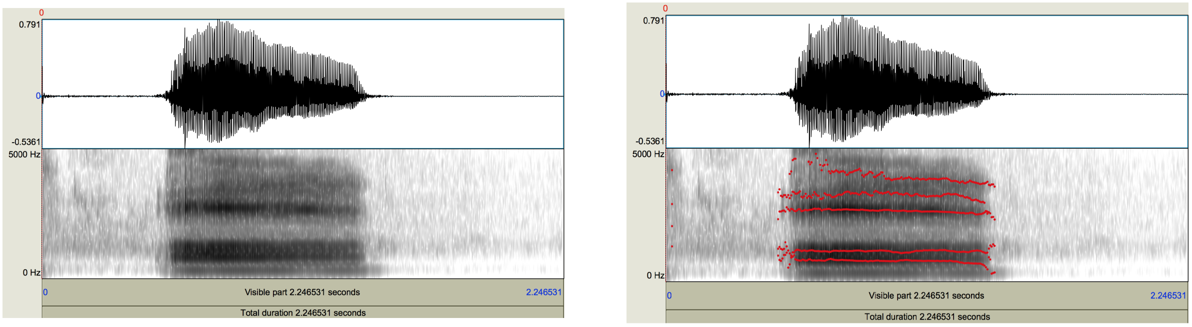


Figure 1.1: An Example of Forced Alignment (from DASS speaker 025; European American Female from Tennessee). Blue line indicates the pitch of the sentence. Phone-level transcriptions are in ARPABET symbols with stress markers.

nasal merger in Southern US English. Forced alignment, which is used in Automatic Speech Recognition (ASR) programs, is the process of automatically generating phone-level transcriptions from orthographic transcriptions (sentence-level transcriptions) of audio; using it, individual phones are aligned to segments of the audio recording. Figure 1.1[3] shows the forced alignment for one utterance of the sentence *I've seen plenty of them.*

Traditionally, two approaches are used in the identification of prenasal merger. First, researchers can manually listen to each vowel uttered by a speaker and check whether the prenasal merger is present. However, this process is laborious and subject to disagreement between listeners. As such, the research community also studies prenasal merger by acoustic analysis, the steps of which include performing forced alignment on the text and analyzing

[3] In the caption of this figure, DASS stands for "the Digital Archive of Southern Speech", which is the audio corpus used in this thesis. We cover some essential details of DASS later in this section. A more detailed introduction of DASS can be found in Section 2.2.



(a)

(b)

Figure 1.2: Two Spectrograms of the Author's Vowel AA; One Without Formants labeled (Left) and the Other With Formants labeled (Right).

the automatically extracted formant values in each of the aligned vowels. Vowel formants are characteristic overtones in the speech signal. Formants can be visualized in an audio spectrogram and two formant values, F1 and F2, represent the height and frontness of the articulated vowel token. On a spectrogram, formants can be seen as darker bands. For example, Figures 1.2a and 1.2b show two audio spectrograms of the vowel /ɑ/ (like the vowel in the word *lot*) with and without labeled formants, processed through the software Praat ([4]). The process of obtaining the formant values of a vowel token is called formant extraction.

Forced alignment requires the use of a pronunciation dictionary, and researchers commonly choose the CMU Pronunciation dictionary,[4] which represents standard American English but does not in particular account for dialectal patterns in Southern US English. Therefore, because a pronunciation dictionary is required and a standard dictionary is commonly used, if the second approach above is taken, the forced aligner cannot choose IH for certain vowel tokens where the prenasal merger is present (e.g., in word *pen*), even if these vowel tokens have acoustic qualities most similar to IH. The forced aligner with an

[4] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

unmodified pronunciation dictionary must transcribe these tokens as EH even though their acoustic qualities are more in conformity with IH. This in turn makes the aggregated acoustic qualities of the prenasal EH tokens (PEN) and prenasal IH tokens (PIN)[5] similar to each other. Specifically, the formant values of PEN obtained using unmodified forced alignment are similar to those of PIN. This similarity, represented by a low Pillai score (which is a result of a MANOVA test)[6] and a low Euclidean distance between the centroids of PEN and PIN clusters, is what phoneticians commonly use in order to identify the prenasal merger.[7]

In this thesis, Kaldi and the Montreal Forced aligner (MFA) are used to perform forced alignment.[8] To facilitate the process of formant extraction on a large corpus, we use a Python wrapper of the software Praat([4]. called FAVE([12]). to automatically extract formant values for the audio files in the corpus.[9]

We hypothesize that if we use a modified pronunciation dictionary which allows the forced aligner to choose between EH and IH in prenasal vowel utterances, then the forced aligner will reclassify some tokens from PEN to PIN (because what the speaker actually utters sounds closer to IH instead of EH).[10] The PEN-to-PIN reclassified tokens should be tokens where the merger occurs for that speaker. We hypothesize that **forced alignment based on a suitably modified pronunciation dictionary can be used to identify prenasal**

[5] Hereafter in this thesis, unless specified otherwise, we use PEN and PIN to refer to the labeled prenasal EH and IH tokens from the forced aligner, respectively. Similarly, we use PET and PIT to refer to transcribed EH and IH tokens in other (i.e. non-nasal) contexts from the forced aligner.

[6] Examples using the Pillai Score to study vowel mergers and splits include [5], [6], [7], and [8].

[7] For instance, see [9].

[8] Kaldi ([10]) is an open-source software developed and maintained by the Center for Language and Speech Processing at Johns Hopkins University. One important step in Kaldi's training pipeline for GMM-HMM models is to align each phone into its corresponding timestamps, which provides users the possibility of directly using Kaldi for forced alignment. Montreal Forced Aligner (MFA), is a Kaldi-based aligner (see [11]). It is a user-friendly Python-wrapper for Kaldi's function of forced alignment. In this work, both Kaldi and MFA are used and the same backend ensures the agreement between their results.

[9] This method of automatically extracting formant values is reliable and commmonly used among sociolinguists and phoneticians. For a comparison between automatically extracted formant values and hand-measured formant values, see [13].

[10] The reliability of this approach has been evidenced by [14], [15], among others, which will be introduced in Section 3.1 in detail. The documentation of this phenomenon can be found in [1], where the author says under this merger "*pen* becomes a homophone of *pin*".

merger, and more specifically, the proportion of pen-to-pin reclassified tokens represents the degree of the merger. We also consider the reverse possibility of PIN tokens merged into PEN, which, although not a characteristic of prenasal merger in Southern US English, could also result in an overlap between PIN and PEN clusters. The proportions of PEN-to-PIN reclassification should be significantly higher than PIN-to-PEN reclassification.

Our method used here is similar to methods used in other studies, including [16], [14], and [15], all of which also use reclassification rates from forced alignment with a suitably customized pronunciation dictionary to identify sociolinguistic patterns. For example, [14] studies the g-dropping (as in the word *walkin'*) by creating two variants – IHN and IHNG – for these words in the pronunciation dictionary and letting the forced aligner choose how frequently the g-dropping happens (that is, the forced aligner chooses IHN instead of IHNG) for selected speakers. Different from these previous studies, we use a new method to verify our results. This method is based on analyses on vowel formant values. Detailed introductions on our methodology can be found in Sections 1.2 and 3.2.

The Digital Archive of Southern Speech (DASS)[11] is selected as the speech corpus to investigate how forced alignment could be used to identify and study the prenasal merger. DASS is an audio corpus featuring interviews conducted from 1968 to 1983, with speech from 30 female and 34 male Southern speakers, totaling 372 hours of audio data. A more detailed introduction on DASS and our processing pipeline can be found in Section 2.2 of the thesis.

1.2 Experiments and Results

In order to verify our research hypotheses, a crucial step is to check whether the PEN-to-PIN reclassified tokens actually correspond to IH. Since manually checking by listening is

[11] A general introduction on DASS can be found in [17]. More recent work on DASS can be found in [18].

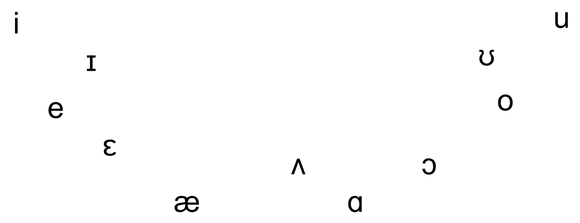


Figure 1.3: American English Vowel Space without any shifts. Revised from [19], where the author introduces acoustic characteristics of the vowel systems of six regional varieties of American English. In this figure, we remove the arrows that indicate regional shifts. Only monophthongs are shown.

not practical in our project, we propose that analyses can be conducted on formant values of PEN and PIN clusters obtained using both the unmodified and modified pronunciation dictionaries. Based on the common understanding of American English vowel space (see Figure 1.3), if the modified forced alignment correctly re-classifies PEN instances as PIN, the acoustic qualities of these instances should be closer to IH (including closer to IH as it appears in PIN and PIT) than to EH (including PEN and PET) in the vowel formant space, and this should be manifested as an increase of separation between the clusters of PIN and PEN, and a decrease of separation between the clusters of PIN and PIT and the clusters of PET and PEN.

In our experiments, we performed forced alignment twice, first with the unmodified pronunciation dictionary, and then with a modified pronunciation dictionary as we have specified above. After performing the alignments, we found that because many PEN tokens were re-classified to PIN after the modification, the centroid of the PEN cluster moved down and right (an increase of F1 and decrease of F2), and because the PIN cluster then contained more tokens (though some PIN instances were re-classified to PEN tokens), its centroid also changed by moving slightly down and right. The research community commonly uses the Pillai score and the Euclidean distance between centroids of vowel clusters to evaluate the overlap and separation between two clusters. We hypothesized that there should be an increase in both the Pillai Score and the Euclidean distance between the centroids of PEN and PIN clusters, and we did in fact see an increase of separation between clusters of PEN and PIN tokens. We did not change any words in the categories of PIT and PET, which are used as controls for comparing with PIN and PEN, respectively. The significant changes are only in the categories of PIN-PEN and PEN-PET. We also perform t-tests on F1 and F2 values of PIN, PIT, PEN, and PET clusters for each speaker before and after the modification in order to examine whether there is a significant change in each of the four distributions. We also performed t-tests on F1 and F2 values of PIN, PIT, PEN, and PET clusters for each speaker before and after the modification in order to see whether there is an significant change. In our results, 48 and 45 speakers had a significant change in the F1 and F2 values of PEN, respectively. 24 and 26 speakers had shown a significant change in the F1 and F2 values of PIN, respectively. None of the speakers showed a significant change in the F1 and F2 values of PIT or PET, which was to be expected because there was no modification on these words in the pronunciation dictionary.

We also investigated whether forced alignment results, especially the proportions of PEN-PIN reclassification, could be directly used to identify the degree of prenasal merger. Specifically, we checked whether there is a correlation between 1) the proportions of PEN-to-PIN re-classification (dubbed as the " PEN-PIN Reclassification Rate", or PEN-PIN RR) for a speaker according to modified forced alignment, and 2) the degree of the speakers prenasal merger according to the Pillai score and Euclidean Distance between clusters of PEN and PIN tokens.

The Pearson Correlation Coefficient and the accompanying p-values between statistics of formant values and Reclassification Rates (including PEN-PIN RR and PIN-PEN RR) are used in checking whether there is such a correlation. Our results showed that a speakers PEN-PIN RR was moderately correlated with the Pillai score between the speakers PIN and PEN clusters (R = -0.506, $p < 0.0001$) and the Euclidean distance between the centroids of these two clusters (R = -0.506, $p < 0.0001$) according to the unmodified pronunciation dictionary. Similarly, pen-pin RR was moderately correlated with the change of the Euclidean distance between the centroids of these two clusters (R = -0.512, $p < 0.0001$). However, pen-pin RR was not as strongly correlated with the change of the Pillai score between pin and pen (R = 0.42, $p = 0.0006$).

In addition to our findings above, we were also interested in whether pen-pin RR tracked other values, including the F1, F2 values of the original pin and pen clusters centroids, and the changes thereof. Our results showed that there are strong correlations between pen-pin RR and the F1 value of pen (R = -0.71, $p < 0.0001$). And pen-pin RR was strongly correlated with the changes of F1 and F2 values of pen (R = 0.73 and -0.74, respectively. In both tests, $p < 0.0001$). We also found a weak but significant correlation between a words pen-pin Reclassification Rate among all speakers and the average F2 pen value of the speaker (R = 0.25, $p = 0.0442$).

We also examined pin-pen RR and studied whether and how it tracked any other statistics of formant values. For a majority of the speakers, the pen-pin RR is higher than the pin-pen RR for the speaker. There were 15 speakers whose pin-pen RRs were higher than their pen-pin RRs. Besides the noise in the audio, a factor that likely contributes the most to the abnormal outcome was that the average F1 and F2 values of originally classified pin tokens of these speakers diverged from the majority, which means that the acoustic space for each of these outlier speakers was likely to be different from the standard one. Consequently, when we used a pre-trained acoustic model trained on standard American English in forced alignment, there were likely to be more inaccuracies among the vowel tokens of these speakers because the model is trained on clean data without many outliers in terms of speech variations. Specifically, we found that a high pin-pen RR is correlated with a high F1 value (R = 0.466, $p = 0.0001$) and a low F2 value (R= -0.45, $p = 0.0002$) of pin tokens.

In addition to using the Pillai Score and Euclidean Distance as indicators of the increase of separation between the cluster of PEN and PIN tokens, we also apply K-means clustering to indicate that the modified forced alignment provides better phonetic transcriptions of PIN and PEN tokens.

If a ground truth existed for a given data set (that is, vowel labels could be correctly attached to each instance), then the accuracy of the phone transcription using a given pronunciation dictionary could be evaluated by comparing it with the ground truth. In reality, however, there are no ground truth labels for IH and EH tokens, so we performed K-means clustering (K = 2) on these tokens (including PIN, PIT, PEN, PET tokens), which gave us two clusters – one higher and fronter (low F1, high F2 values), and the other lower and backer (high F1, low F2 values) in the vowel space. We performed post-hoc visual inspections and removed 8 out of 64 speakers whose K-means clustering results were not following this pattern because in the vowel plots of these speakers, the locations of IH and EH do not follow the patterns defined by the common understanding of American English vowels. For instance, for some speakers, one cluster has high F1 and high F2 values and the other has low F1 and low F2 values, which was an indicator that the speakers distribution of vowel measurements was skewed and therefore should be abandoned.

In the K-means clustering results after filtering, the first cluster (the higher cluster), based on the vowel space indicated in Figure 1.3, was labeled IH (including PIT and PIN), and the second cluster was labeled EH (including PET and PEN). Our modified forced alignment also generated two clusters, and we checked how many tokens were correctly identified in the forced alignment results relative to the two clusters produced by K-means clustering. We showed that when the modified pronunciation dictionary was used, there were improvements in both precision and recall for identifying IH relative to K-means clustering results. Detailed discussions on this topic can be found in Section 3.5 of the thesis.

Forced alignment can also be used to produce acoustic models to map phone labels to the processed audio features for the purpose of phone classification. We generated acoustic models and use them to examine how the prenasal allophones of IH and EH are different from each other, especially where in the duration of an utterance these two vowels differ the most. We treated prenasal IH and EH as two separate phones — IHN and EHN — in the pronunciation dictionary, which corresponded to the contexts of the previously mentioned PIN and PEN tokens. Next, by leveraging the forced aligner (specifically, its ASR backend), we trained customized acoustic models with and without modifications on pronunciation dictionary in accordance with the prenasal merger. After training, each of the trained acoustic model had 3 Hidden Markov States (HMMs) and each state could be represented by a mixture of Gaussian distributions. To evaluate the difference between two different Gaussian Mixture Models, Kullback-Leibler divergence (KL-divergence) was used. Our results suggest that among the 3 Markov states (each of which is a GMM model), IHN and EHN differ the most in the last 2 states after the modification, whereas the first state of IHN and EHN stay roughly the same, which is also where IHN and EHN are most similar with each other. This indicates that after customized training with the modified pronunciation dictionary, *the acoustic models* used for distinguishing between IHN and EHN mainly differ on the last HMM states. Detailed comparisons on the values of KL-divergence mentioned above can be found in Chapter 4 of the thesis.

1.3 Outline of the Thesis

In the following parts of this thesis, Chapter 2 introduces more background information, including descriptions on the prenasal merger, the DASS corpus, and the fundamentals of forced alignment. Chapter 3 discusses our experiments and the results based on modifying the pronunciation dictionary in forced alignment. Chapter 4 gives a more detailed introduction on GMM models and HMM models, and shows how to compare the KL-divergence between different acoustic models. Results of the comparisons are used in identifying where the acoustic models for the prenasal allophones of IH and EH differ the most. Chapter 5 provides our conclusion and suggestions on how this work could be extended. A set of appendices provides more detailed facts and figures related to our results.

Chapter 2

Background

2.1 Characteristics of Southern Speech

2.1.1 Relevant concepts in Phonetics and Phonology

In this section, we introduce some background concepts of phonetics and phonology that are crucial to discussions of sound variations and distinct characteristics of Southern speech.

Place of articulation refers to where a sound is made. When humans talk, the air from the lung passes between two small muscular folds. These folds are called the *vocal folds*, and the air passage above the larynx is called the *vocal tract. Articulators* are parts of the vocal tract that can be used to form sounds. Types of articulators include tongues and lips. In the production of vowel sounds, the passage of the airstream is relatively unobsrtucted. Figure 2.1 is an image of different positions of vocal organs in the productions of different sounds.

We can categorize different types of vowels by the height of the body of the tongue, the front-back position of the tongue, and the degree of lip rounding. We mainly discuss the first two features here. When producing front vowels, the highest point of the tongue is in the front of the mouth, and the tongue is fairly close to the roof of the mouth, for example, the word *heed. Hid* is a little lower than *heed*, and *head* is lower than *hid.* When producing

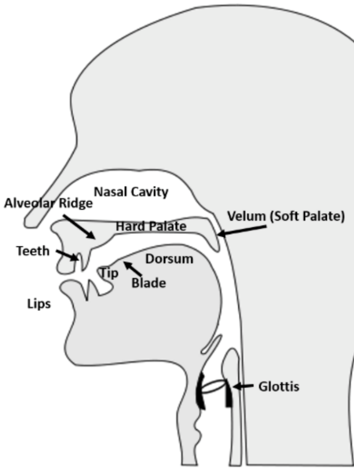


Figure 2.1: Places of Articulation. From [20].

back vowels, tongue is close to the back of the surface of the vocal track. For example, /u/ in *food* and /ɑ/ in *father* are both back vowels. The difference is that /u/ is a high back vowel and /ɑ/ is a low back vowel. /ɪ/ is a high front vowel whereas /ɛ/ is a mid front vowel, as in words *hid* and *head*, respectively. Speakers can get some impression of tongue height and feel the difference between front and back vowels by comparing different pairs of words. Figure 1.3 in Chapter 1 visualizes the positions of standard American English vowels in IPA symbols.

Manners of articulation are basic ways in which articulatory gestures can be accomplished, including stop, nasal, fricative, approximant and lateral. They can be used to represent different consonants. A nasal consonant is a sound where the air is stopped in the oral cavity but the velum is lowered so that the air can go out through the nasal cavity. Nasal consonants include /m/, /n/, and /ŋ/, which are represented as M, N, and NG in ARPABET.

Quantitative measurements on these features can be achieved by measuring the formants of vowels. To understand the definition of a formant, we first need to understand the definition of overtone. A vowel contains a number of different pitches simultaneously. There is a pitch at which the vowel is actually spoken by the vibrating vocal folds, which is called the fundamental frequency F0. And there are also *overtone* pitches that depend on the shape of the resonating cavities of the vocal tract, which give the vowel its other distinctive qualities.[1] These overtones, which are also called formants, are helpful for us to distinguish between different vowels. The lowest formant (F1) reversely corresponds to the height of the vowel. The lower the F1 value is, the higher the position of the tongue. Formant two (F2) reversely corresponds to the backness of the vowel. The higher the F2 value is, the more front the position of the tongue is.[2] On a spectrogram, formants one and two show up as darker horizontal lines. Figure 2.2 shows the formant of vowel OW from DASS speaker 472. To extract the formant values of a vowel, the sound processing software Praat ([4]) is usually used. To extract a large amount of vowel formant values, Praat scripting or a software wrapper is often used together with the FAVE (Forced Alignment and Vowel Extraction) Program Suite.([12]) As the name indicates, FAVE has two distinct modules for two different tasks – Forced Alignment and Formant Extraction, which are distinct but closely connected in the research of phonetics. This project only uses the Formant Extraction function of FAVE. In Section 3.2.1 of this thesis, we explain the reason for doing so.

2.1.2 The Prenasal Merger

One method for researchers to study dialectal features is to examine formant values in vowel space and study how they change in specific accents. [3] lists 22 distinct characteristics of Southern speech in 7 different categories. These features reflect how Southern accents are differentiated and deviate from the standard American accent (although some of them are not limited to the region).

[1] This is discussed in detail in [20], p.187.

[2] For a more detailed introduction, see [20], Section 8.4.

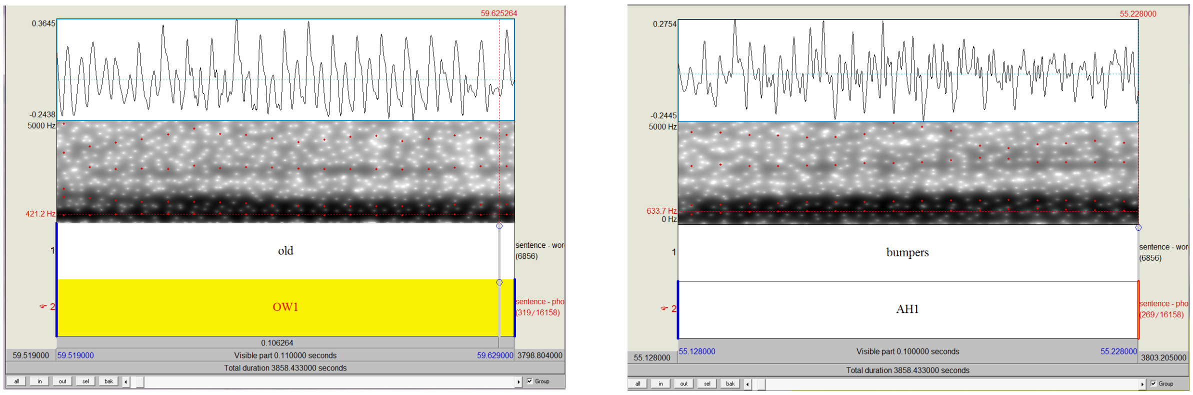


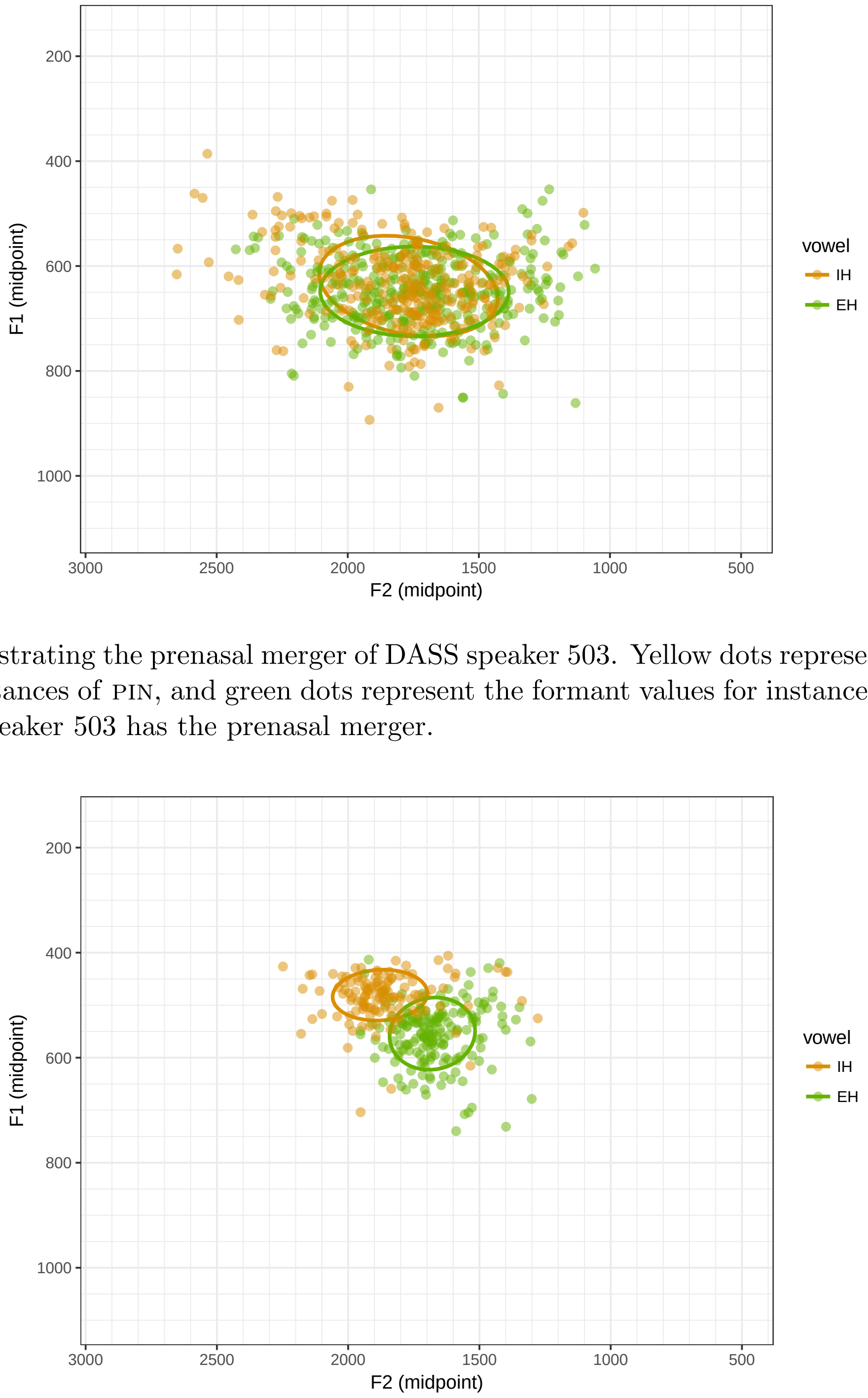
Figure 2.2: Formant values of OW and AH from DASS speaker 472 (European American; M; Birth Year 1904) shown in Praat. From an impressionistic perspective, he has a very low voice, which is manifested as the close distance between F1 and F2 in the spectrograms above.

One approach for sociolinguists to study the features of Southern speech is by measuring the F1 and F2 values and conduct analyses on them. As we have stated before, F1 and F2 values of a vowel represent the height and frontness of the vowel. Figure 2.3a is a plot representing a part of the vowel measurements of DASS speaker 503.[3] The plot represents the F1, F2 values of prenasal allophones of IH and EH (i.e. tokens of PIN and PEN). The clusters of PIN and PEN have a high degree of overlap. The mean F1, F2 values of PIN and PEN are also close, which indicates that the places of articulation of PIN and PEN are also close. In contrast, Figure 2.3b is a plot representing a part of the vowel measurements of DASS speaker 657X.[4] Similarly, the plot also represents the F1, F2 values of prenasal allophones of IH and EH (i.e. tokens of PIN and PEN). However, the clusters of PIN and PEN have a low degree of overlap, which indicates that the places of articulation of PIN and PEN are not close to each other. We can claim that the prenasal merger is not present in his speech.[5]

[3] European American Male from Tennessee; Born in 1893; Retrieved from [21] on July 6th, 2019.

[4] European American Male from Louisiana; Born in 1897; Retrieved from [21] on July 6th, 2019.

[5] From the perspective of articulatory phonetics, the reason for the presence of the prenasal merger is



(a) A plot illustrating the prenasal merger of DASS speaker 503. Yellow dots represent the formant values for instances of PIN, and green dots represent the formant values for instances of PEN. This plot shows speaker 503 has the prenasal merger.

(b) A plot illustrating the absence of prenasal merger for DASS speaker 657X. Yellow dots represent the formant values for instances of PIN, and green dots represent the formant values for instances of PEN. This plot shows the prenasal merger is not present for the speaker.

Figure 2.3: Plots of DASS speakers 503 and 657X.

2.2 DASS Corpus and Previous Studies

DASS ([17], [18]) is a part of LAGS (The Linguistic Atlas of the Gulf States, see [25]), which contains data from 1121 speakers. Reel-to-reel cassette audios have been digitized and maintained by the Linguistic Atlas Project at the University of Georgia. These speakers are representative of both rural and urban communities and a variety of social and age backgrounds.

DASS is a subset of 64 LAGS interviews selected by Lee Pederson to provide a balanced representation of speaker types from different regional sectors in the American South. For each of the sectors, researchers selected three European Americans and one African American. In total, there are 30 female and 34 males, ranging from 15 to 90 years old. Duration of interviews average about 5.75 hours, range from 2.5 hours to 10 hours. Figure 2.4 shows the geographical distributions of the 64 speakers.

Speech data in DASS is semi-spontaneous. There are some targeted words for the interview, but many conversations are free flowing about the interviewee's lives so the corpus covers many words and topics. Due to its span of years of recording and a variety of recording conditions (including noise, background music, etc.), audio qualities of DASS reels also vary. Sometimes there are also overlaps between the interviewer and the interviewee. All these factors pose unique challenges for ASR, forced alignment, and formant extraction on DASS.

[18] has documented the recent work on DASS in detail. 48 University of Georgia undergraduate transcribers have contributed to the orthographic transcriptions of DASS. There are in-house protocols for transcription and an in-house dictionary for non-standard words like *uh-huh*, *gonna*, *wanna*). The transcription quality is assured by a four-listen system —

that when a vowel is articulated before a nasal consonant, the nasal consonant introduces another cavity — nasal cavity — in addition to the oral cavity, which changes the oral cavity shape. Relevant discussions can be found in [22], [23], and [24].

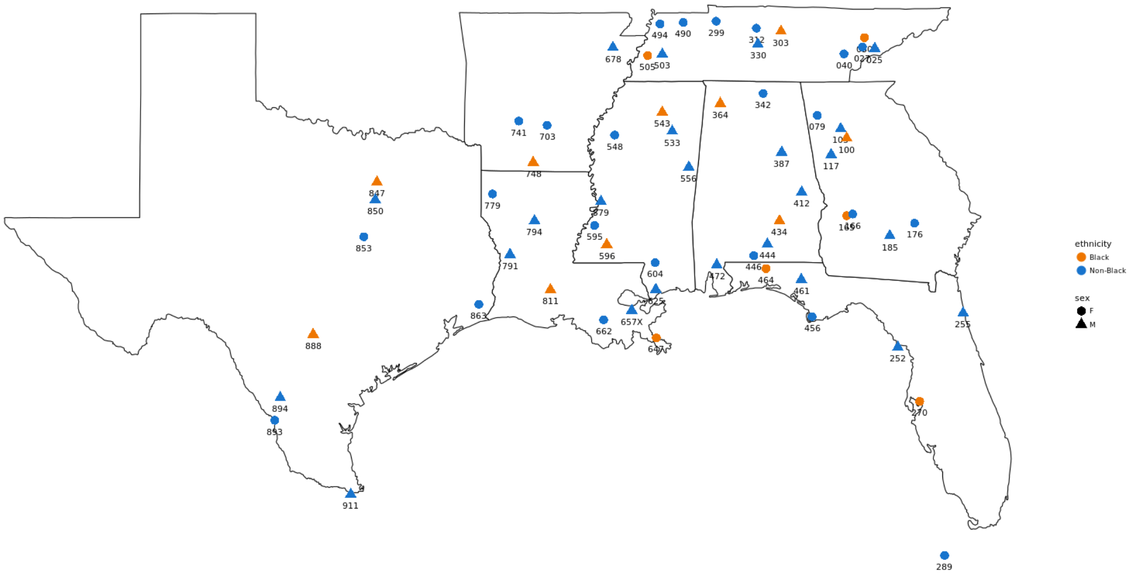


Figure 2.4: Distributions of the 64 speakers in DASS. Shapes indicate the sex of each speaker (Circle: Female; Triangle: Male); Colors indicate the ethnicity of the speaker. (Orange: African American; Blue: European American). From [21]

undergraduate transcribers listen to the audio twice, and graduate assistants check the third time and make corrections if necessary. Afterwards, undergraduate assistants check the transcripts with the help from a spellchecker. Noise reduction tools, such as SoX[6] and Audacity ([26]) are also used to obtain audios with improved quality.[7] The three main methods for noise reduction include applying a low-pass-filter, using Audacity, and using SoX. The results show that in the three-way comparison, both low-pass-filtered and Audacity-noise-removed reduce outliers in vowel formant plots. In this work, we use low-pass-filtered audios.

[6] Available at http://sox.sourceforge.net/.

[7] [27] documents the work of noise removal in detail.

Table 2.1: Part of the CMU Dictionary.

DATA	D EY1 T AH0
DATA(1)	D AE1 T AH0
DATABASE	D EY1 T AH0 B EY2 S
DATABASE(1)	D AE1 T AH0 B EY2 S
DATABASES	D EY1 T AH0 B EY2 S IH0 Z
DATABASES(1)	D AE1 T AH0 B EY2 S IH0 Z
DATE	D EY1 T
DATE'S	D EY1 T S
DATED	D EY1 T IH0 D

2.3 Pronunciation Dictionaries

In order to perform forced alignment on audio files with orthographic transcriptions, a pronunciation dictionary is needed. Common English pronunciation dictionaries[8] use ARPABET symbols[9] to encode a word's pronunciation at the phone level. Table 2.1 shows a part of the CMU Pronunciation Dictionary. The pronunciation for each word (shown in the left column) is listed in the right column using ARPABET symbols. The number after each vowel indicates the vowel's stress marker (1: primary stress; 2: second stress; 0: no stress). Words with multiple pronunciations (for example, the word *data*) are listed on multiple lines with increasing indices, and these can be used to select the best possible pronunciation in the process of forced alignment. In the next section, we introduce some technical details on how variations in pronunciation are identified in the process of ASR and Forced alignment.

2.4 Automatic Speech Recognition and Forced Alignment

In order for an ASR system to distinguish between different phones, it needs to train an acoustic model and use it to map from acoustic features to phone labels. Consequently, forced

[8] For instance, the CMU Pronunciation Dictionary on the website http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[9] More details on ARPABET can be found in Appendix A and Table A.1.

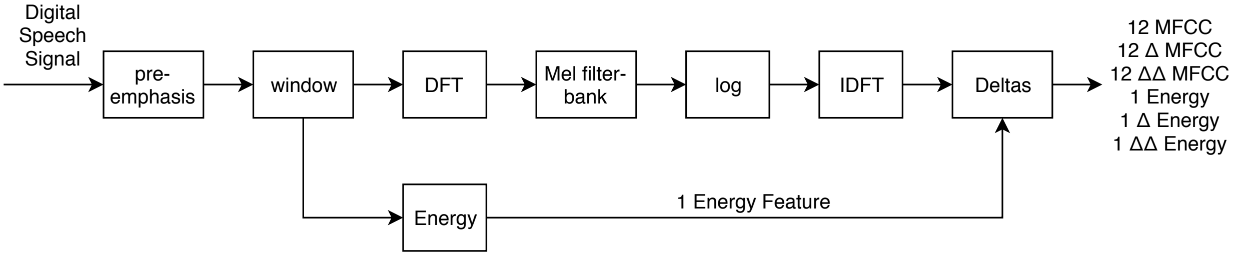


Figure 2.5: Steps for Extracting Features for ASR. From [28], Section 9.3.

alignment is one of the main steps of the ASR training process. This chapter reviews the common pipeline of ASR systems based on Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), especially how forced alignment implemented in the process.[10]

2.4.1 Feature Extraction

ASR systems use signal processing tools to extract acoustic features from speech waveforms and build acoustic models to map speech features to phone labels. Figure 2.5 illustrates the steps needed to extract features for ASR. The first step is to convert analog representations (from the recording device) to digital representations. Since an audio signal is continuous, certain “windows” (intervals) are needed to detect and extract these features. Hamming windows, which avoid the audio being cut abruptly at boundaries, are commonly used. Each window usually lasts 25ms and the interval between each window is 10ms. To extract features from each window, Fast Fourier Transformation (FFT) is usually used. The results of FFT show the amount of energy at each frequency band. Outputs of frequencies from FFT are transformed to a mel scale. A mel is a unit of pitch defined so that pairs of sounds which

[10] There are more advanced ASR systems based on deep neural networks, and they produce state-of-the-art results. But the current mainstream technique used in forced alignment is still GMM-HMM-based. Moreover, the GMM-HMM acoustic models are more interpretable.

are perceptually equidistant in pitch are separated by an equal number of mels.[11] From the Mel spectrum, the next step is to compute cepstrum, which is calculated by first taking the logarithm of the Mel spectrum and then visualizing the log spectrum as if itself were a wave form.

To extract MFCC (Mel Frequency Cepstral Coefficents) features, we extract the first 12 cepstral features and also add energy, which can be calculated by summing over time of the power of the sample in a frame. Energy is helpful in distinguishing vowels and stops, because vowels have more energy than stops. So far, there are 13 features in total. Many ASR systems use 39-dimensional feature vectors, corresponding to twelve MFCCs, plus energy, and their first and second derivatives (delta and delta-delta features).

The next step is to map the 39-dimensional MFCC features in each window to labels that are meaningful to ASR. These labels could be a phone, or a subphonemic state.[12] One way to build this mapping relation, as we will introduce in the next section, is by Gaussian Mixture Models. Since these windows are sequential, a sequence classifier is also needed in the decoding process. Hidden Markov Models (HMMs) are usually used as the sequence classifier. In the next section, we give a brief introduction to GMMs and HMMs as well as their implementations in ASR.

[11] See [29]. There are also other features that are used in forced alignment and speech recognition. For instance, PLP features is used by The Penn Phonetics Lab Forced Aligner [30]. But Kaldi and MFA use MFCC features. (Mel Frequency Cepstral Coefficents) Hence MFCC features are discussed here. The conversion between the mel frequency m and the raw acoustic frequency f can be computed as follows:

$$mel(f) = 1127 \ln(1 + \frac{f}{700}) \tag{2.1}$$

[12] A subphonemic state models part of a phone. If subphonemic states are used, a phone is divided into multiple distinct parts, and researchers treat these parts as distinct phones, and train different acoustic models for them. For instance, we can use IH_B to denote the beginning part of the phone IH and train a separate model for it.

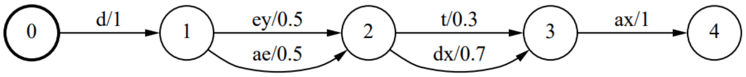


Figure 2.6: Lexicon FST for the word *data*.

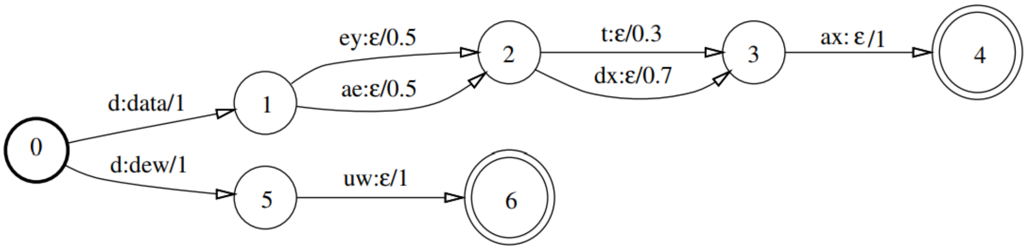


Figure 2.7: Lexicon FST for the words *data* and *dew*. From [31].

2.4.2 Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Finite State Transducer (FST)

As its name indicates, a Gaussian Mixture Model (GMM) is a mixture or combination of multiple Gaussian distributions. Using the concepts introduced in the last section, a GMM can be used to characterize the feature of a specific 25-ms window. However, speech is a discrete combination of these windows. Therefore, methods that characterize time-series transitions need to be used. That is where Hidden Markov Models (HMMs) are introduced and used in combination with GMMs in ASR systems. Because the results of GMMs are probabilistic, HMMs are considered to be probabilistic as well.

In order to use probabilistic HMMs to train and perform forced alignment based on orthographic transcriptions and the pronunciation dictionary, another type of inputs are needed – weighted Finite State Transducers (FSTs). By maniputating the pronunciation dictionary, we change the Lexicon FSTs. Figures 2.6 and 2.7 are two examples of FST, from [31]. FST is able to handle a word with multiple pronunciations by adding connections between different states. In the forced alignment process, the aligner uses the Viterbi algorithm to find the best path along the FSTs of the words in a sentence in combination with trained GMM-HMM models.[13] Montreal Forced Aligner is shipped with English acoustic models trained on LibriSpeech ([32]) and it is used in some of our experiments as well (this is specified in the next chapter).

A GMM-HMM model also needs to be trained before it is used in finding the best alignment. Several things are needed in order to train a model. First, Kaldi initiates a GMM, which includes initialization of the probability density functions (pdfs) for each phoneme. Lexicon FSTs are also used in this step. If there are words with multiple pronunciations, Kaldi randomly selects a path in the FST of the sentence.[14] Based on this path, Kaldi initiates an equal alignment of the phones for the sentence, and then it evaluates the mean and covariance of each GMM based through the process of Viterbi training. After a certain number of iterations of training (usually 10), the forced aligner re-aligns the text into phonemes and use the updated alignment to update the parameters of GMM-HMM models. In the script for training a monophone acoustic model (used in Chapter 4), the default setting is 40 iterations, which is also what we use in Chapter 4.

[13] In real-world application of ASR, Lexicon FSTs and GMM-HMM models are usually also combined with FSTs of phonetic contexts and Language Models. In our work, we use monophone models for forced alignment, which are context-independent and hence makes the phonetic context trivial. Language models provide likelihood for a certain word based on nearby words. In forced alignment, language model is also trivial because the forced aligner already knows what the sentence is and does not need the linguistic context information to select the best word sequence.

[14] This is where Kaldi is different from HTK ([33]), and might potentially contribute to the difference between results of the forced aligners based on them. In the forced alignment process, HTK selects the *first* pronunciation of each word in the pronunciation dictionary, instead of adding randomness in the process. See [33], p.32.

Chapter 3

The Effects of Modifying Pronunciation Dictionaries on Forced Alignment and Formant Extraction

3.1 Previous work on using forced alignment to model accent features

Researchers have used the technique of forced alignment in order to study regional variations of speech (e.g., [16], [14], [34], and [15]). By modifying the pronunciation dictionary in accordance with the studied speech pattern, researchers are able to identify the variations of pronunciation at the phoneme level. This section reviews some of the previous work based on this approach.

[14] studies g-dropping in American English (e.g., in the word *walkin'*) and compares the accuracy of identifying the phenomenon between native and non-native English speakers. In the pronunciation dictionary, the authors explicitly build pronunciation variations for words

ending with *ing*. The variations include IHN and IHNG, which are treated as two different unitary phonemes in the pronunciation dictionary. In the process of forced alignment, the aligner[1] automatically determines whether IHN or IHNG is present. For instance, the word *coming* has only one pronunciation – **K AH M IH NG** in the CMU Pronunciation Dictionary:[2] and this is expanded to **K AH M IHN** and **K AH M IHNG** in the new pronunciation dictionary. The main corpus used in [14] is the Buckeye Corpus ([35]), which includes manually transcribed phone-level labels. The methods to determine whether the forced aligner correctly distinguishes between IHN and IHNG is to check whether they are in agreement with human judgments. [14] also tests the presence of g-dropping across different speakers on the SCOTUS corpus,[3] and shows that different judges have different g-dropping rates.[4]

[15] considers three sociolinguistic variables in British English – (th)-fronting, (td)-deletion, and (h)-dropping.[5] By using the same method of expanding the pronunciation dictionary, the forced aligner is able to determine whether these variations are present. The author compares the results from the forced aligner with those from human transcribers, and shows that the results from the forced aligner has a higher degree of accuracy in identifying these sociolinguistic variables. The accuracy rates are 85.5%, 79.7%, and 71% for (h)-dropping, (th)-fronting, and (td)-deletion respectively.

The same method can also be used in studying other languages. [34] studies the dialectal

[1] In their work, the aligner is the Peen Phonetics Lab Forced Aligner. See [30]. The aligner is available at http://www.ling.upenn.edu/phonetics/p2fa/).

[2] Here, the version without stress marker is used.

[3] The SCOTUS corpus includes more than 9,000 hours of recordings of the U.S. Supreme Court's oral arguments since the 1950s. A helpful article on the corpus is [30].

[4] [14] also uses KL-divergence to compare the distance between the GMM models of IHN and IHNG at each of the 5 HMM states. A more detailed review is given later in chapter 4 of the thesis.

[5] Here are the examples [15] gives for these three variations:

- (th)-fronting: for word *just*, the pronunciation is changed from **J AH1 S T** to **J AH1 S**.
- (td)-deletion: for word *north*, the pronunciation is changed from **N AO1 TH** to **N AO1 F**.
- (h)-dropping: for word *height*, the pronunciation is changed from **H AY1 T** to **AY1 T**.

difference between French spoken in Québec and France and shows the difference of schwa insertion following a word-final consonant cluster.

Although these studies have produced compelling results, there are also some limitations. First, these studies mainly focus on variations in consonants instead of vowels, but sociolinguistic variations are present in vowels as well. Moreover, all of them need manually-labeled phone-level transcriptions to validate the results, which would take a significant amount of efforts to obtain. Most audio corpora do not come with phone-level transcriptions created by listeners. In addition, both [14] and [15] use the Penn Lab Forced Aligner, which uses HTK ([33]) as the backend. The performance of the Penn Lab Forced Aligner has been superseded by the Montreal Forced Aligner (hereafter MFA, see [11]), which is based on Kaldi. Although Kaldi and HTK share some similarities for the implementation of HMM/GMM models, the performance of Kaldi is consistently better than HTK.[6]

In our work, we use the same method of performing forced alignment with a modified pronunciation dictionary in order to study a kind of variation, specifically, the prenasal vowel merger (PIN-PEN merger) in Southern US English. It is a distinctive vowel variation found in the Southern speech. Proportions of changes from PEN to PIN tokens after the re-alignment are taken as an indicator of the degrees of prenasal merger. Moreover, in contrast to the previous studies using the same method (as reviewed above), we verify the results in a new way that does not rely on manually-labeled phonetic transcriptions. Previous studies of the prenasal merger are mainly based on formant analysis, which examines two sets of values. The first is the Pillai score between PEN and PIN clusters, and the second is a set of distances between PEN and PIN clusters, including the Euclidean distance between the centroids of the two clusters, and the differences of F1 and F2 values between PEN and PIN clusters. In our work, we show that there is a correlation between the *proportions* of PEN-to-PIN changes and the degrees of prenasal merger suggested by the metrics based on formant analysis. Previous

[6] For detailed comparisons on HTK and Kaldi and forced aligners built upon them, see [10] and [11].

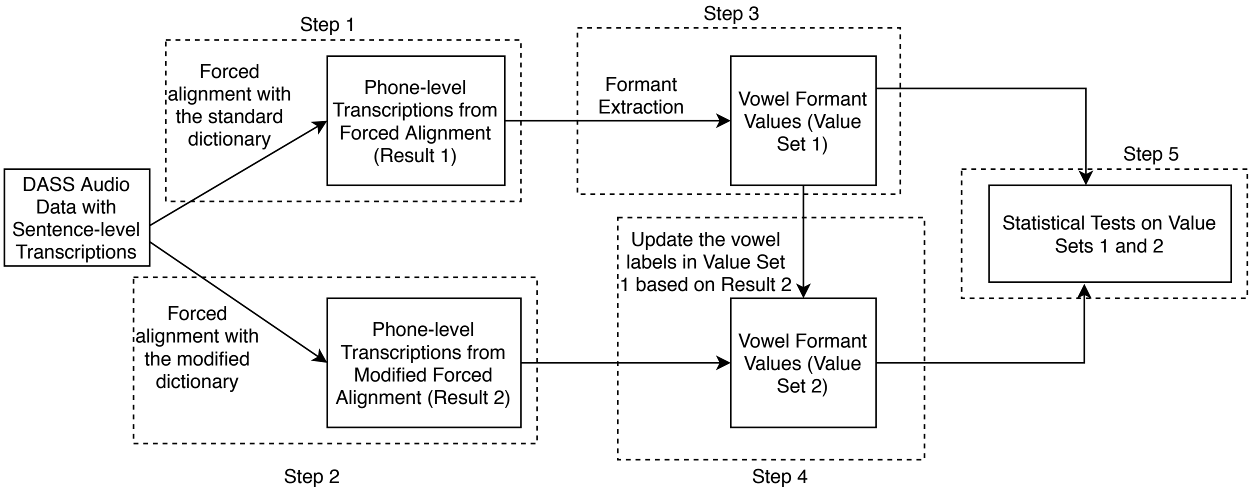

Step 1
Forced alignment with the standard dictionary
Phone-level Transcriptions from Forced Alignment (Result 1)
DASS Audio Data with Sentence-level Transcriptions
Step 3
Formant Extraction
Vowel Formant Values (Value Set 1)
Step 5
Statistical Tests on Value Sets 1 and 2
Update the vowel labels in Value Set 1 based on Result 2
Forced alignment with the modified dictionary
Phone-level Transcriptions from Modified Forced Alignment (Result 2)
Vowel Formant Values (Value Set 2)
Step 2
Step 4

Figure 3.1: Steps for using forced alignment to identify prenasal merger and study whether it is able to provide better phonetic transcriptions.

studies that use forced alignment to study sociolinguistic variations are on the variations of consonants, which could not be studied by using formant values.

3.2 Methodology

3.2.1 Steps of Experiments

In our experiments, some pre-processing steps were needed before the Kaldi-based Montreal Forced Aligner could be used for forced alignment on DASS. We first applied a low-pass filter on all the audio files in DASS. Low-pass filtering only keep the audio frequencies in a certain low range, which is what humans mainly hear and therefore can preserve the quality of the audio file without a significant loss of information. We also removed all punctuation markers. Only apostrophes within a word (as in *it's* and *they'll*) were kept. All transcriptions were also changed to the upper case.

Figure 3.1 shows the steps of our experimental pipeline. After the processing steps, we performed forced alignment on the audio corpus using an unmodified pronunciation dictionary. In order to generate such a pronunciation dictionary, we obtained a list containing all words in the corpus. Next, the CMU Pronunciation Dictionary (the CMU Dict) was used as the default pronunciation dictionary. If a word was in the CMU Dict, then it was considered to be an in-vocabulary word and we directly used its pronunciation(s) provided by the CMU Dict. For words that were not in the CMU Dict (i.e. out-of-vocabulary words), we used an open-source grapheme-to-phoneme toolkit Phonetisaurus to generate their pronunciations.[7] At this point, forced alignment could be performed on DASS, and afterwards, we had phone-level transcriptions and timestamps (beginning and end) for each audio file. The results from this step were called *Result 1*.

Next, we modified the pronunciation dictionary based on the patterns suggested by the prenasal merger. Alternative pronunciations were added. Specifically, words whose transcriptions contained the following patterns on the left-hand side of the arrow were identified, and new pronunciation variants were added by changing EH to IH and IH to EH respectively.[8]

- EH N ⟶ IH N
- EH M ⟶ IH M
- EH NG ⟶ IH NG
- IH N ⟶ EH N
- IH M ⟶ EH M
- IH NG ⟶ EH NG

For example, the word *send* only had one pronunciation S EH N D. After the modification, another pronunciation of S IH N D was added into the pronunciation dictionary. Moreover, the reason we included words whose pronunciations included IH followed by a nasal consonant (as in words *pin*, *tin*, etc.) was that we wanted to show there was not as much PIN-to-PEN reclassification as PEN-to-PIN reclassification.

In total, the pronunciations of 1015 PEN words and 3148 PIN words were modified. The

[7] The reasons we use Phonetisaurus include its fast speed and state-of-the-art results. See [36].

[8] For the purpose of brevity, the stress markers are not present here.

modified pronunciation dictionary was used in another run of forced alignment (*Step 2* in Figure 3.1). In this step, a new version of the phone-level transcriptions was generated by the Montreal Forced Aligner. This new set of phone-level transcriptions was used in the next step for re-assigning phone labels in the formant data. Results from the Montreal Forced Aligner in this step provided another set of timestamps for each phone, and we only kept the phone tokens that differed from the old tokens in less than or equal to 20ms, which included more than 92.3% of all the tokens in Result 1.

The next two steps were to generate two sets of formant values. Step 3 generated the formant values based on Result 1. This process was done automatically by using the formant extraction tool FAVE (Forced Alignment and Vowel Extraction) Program Suite.[9] The default settings of FAVE were used. The inputs of FAVE included an audio file and the phone-level transcriptions of the audio file (Result 1). The result of this step was called Value Set 1 in Figure 3.1.

By examining the Pillai scores between PIN and PEN clusters and the Euclidean distances between the centroids of PIN and PEN clusters in Value Set 1 for each DASS speaker, we could infer the degree of merger for each speaker. A high degree of the prenasal merger is usually evidenced by a low Pillai score and a low Euclidean distance. We use the Pearson Correlation Coefficient to show there is a correlation between the PEN-to-PIN Reclassification Rates (or PEN-PIN RR) and the values of those indicators of the prenasal merger from formant analysis.

Next, in order to generate another set of formant values to compare with Value Set 1, the phone-level transcriptions from Result 2 were used. In Step 4, results after modification were used, and vowel labels were updated.[10] Results are presented in Section 3.3.1 of the

[9] See [12]. Note that although FAVE's name indicates that it does both forced alignment and formant extraction, we only use FAVE's formant extraction function. The reason, which we have stated before, is that FAVE's module of forced alignment is based on HTK ([33]), but HTK has been superceded by more recent toolkits like Kaldi and other deep-learning-based recognizers. Thus the process of forced alignment is performed by Kaldi in this research.

[10] One concern reader might have is whether the formant values from the FAVE should stay the same after this change. As a formant extraction tool, FAVE does leverage some outlier removal and formant prediction

thesis. These updated tokens in Value Set 2 are supposedly where the prenasal merger is present. The vowel labels on the corresponding rows of these tokens were also updated in Value Set 1, which resulted in a new set of formant values – Value Set 2.

The last step was to compare Value Set 1 with Value Set 2. We are interested in whether the clusters of PIN and PEN were changed significantly between these two sets. The hypothesis is that because of the prenasal merger, many tokens that are classified as PEN in Value Set 1 sound like IH, which means that the acoustic qualities of these tokens should also be close to other existing tokens of PIN. Forced alignment based on the modified pronunciation dictionary would presumably identify these "merged" tokens and re-classify them into IH. As a result, after the re-classification, there should be a greater separation between the clusters of PEN and PIN, which can be quantitatively measured by the Pillai score and the Euclidean distance between PIN and PEN clusters.

3.2.2 Euclidean Distance and Pillai Score

The Euclidean Distance between two clusters is usually based on the Euclidean Distance between the *centroids* of these clusters. The centroid of a cluster is where the mean F1 and F2 value of the cluster is.

Pillai score is formally known as the Pillai-Bartlett trace, and it is part of the output of a MANOVA test. Multivariate analysis of variance (MANOVA) tests can characterize variations with respect to more than one dependent variable simultaneously. In recent years, phoneticians have used the Pillai score (notable examples include [9] and [6]) for estimating the extent of overlap between different vowels. A high Pillai score indicates a high degree of separation between two distributions with respect to the dependent variables (in this case, F1 and F2), and vice versa.

methods based on the acoustic data of standard American English. In our work, in order to alleviate the effect of this process, we turned off the options of outlier removal and formant prediction in order to make sure Value Set 2 and Value Set 1 are comparable.

As a result from MANOVA, the Pillai score is supposed to be non-negative. In order to capture the *relative locations* of different vowel clusters (especially where two vowel swap spaces), [6] proposes a metric of "adjusted Pillai score", which is to use visual inspections to make post-hoc adjustments. In our experiments, we made automatic adjustments based on the mean F1 values of the PIN and PEN clusters, which represents the heights of vowels. According to the common notion of American English vowel space (see Figure 1.3), F1 values of PIN tokens are lower than those of PEN tokens. So in order to adjust the Pillai score for the prenasal merger, we compared the mean F1 values of PIN and PEN.

$$M(F1_{\text{PIN}}) > M(F1_{\text{PEN}}) \tag{3.1}$$

If the above inequality holds, then it means that the mean F1 value of PIN is lower than that of PEN and we changed the speaker's PIN-PEN Pillai score to its negative. For instance, Figure 3.2 is a plot of the PIN and PEN tokens of DASS speaker 330.[11] Yellow and green dots represents PIN and PEN tokens, respectively. The ellipses stand for the 95 percent confidence interval for each clusters of data points. The mean F1 value (564.87) of PIN is greater than the mean F1 value (542.60) of PEN, which indicates not only a prenasal merger, but also a swap of vowel spaces of PIN and PEN. In this scenario, the Pillai score between these two clusters should be adjusted to its negative (from 0.016 to -0.016). The same method of adjustment was also applied in calculating the adjusted Pillai scores in other phonetic contexts.

In general, if the Pillai score between PIN and PEN is low from the first run (i.e. Value Set 1), we can claim there is a merger between PIN and PEN. If the Pillai Score from the second run (Value Set 2) is greater that from the first run (Value Set 1), then we claim it indicates that forced alignment by using the modified dictionary is making a better contrast between PIN and PEN. This better contrast is in better conformity of the standard American

[11] European American Male from Tennessee; Born in 1906. Plot is retrieved from [21] on May 16th, 2019.

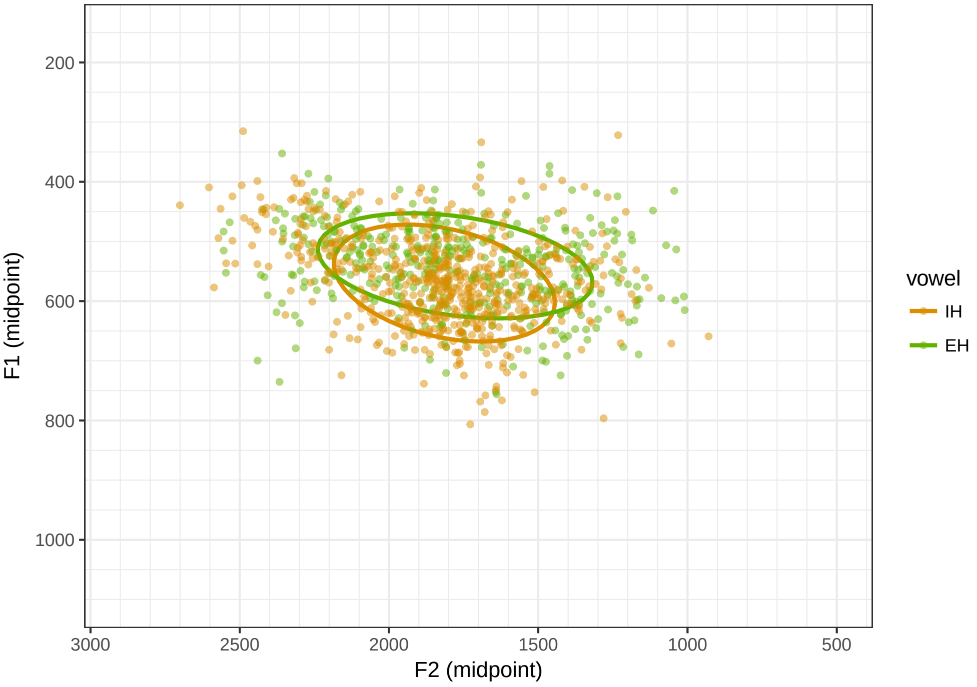


Figure 3.2: Formant values for PIN and PEN tokens of DASS Speaker 330. (Mean F1 value of PIN instances: 564.87; Mean F1 value of PEN instances: 542.60; Pillai Score (unadjusted): 0.016) The ellipses stand for 95 percent confidence intervals. This plot shows an example where the adjusted Pillai score should be used.

English vowel space, which is what we would expect if the prenasal merger is present. So this better contrast constitutes a qualitative indicator of the prenasal merger. Besides, we also calculate the Euclidean distances between the centroids of these clusters in order to examine whether there is a better separation after using the modified pronunciation dictionary. We also perform t-tests on F1 and F2 values of PIN, PIT, PEN, PET clusters for each DASS speaker before and after the modification in order to examine whether there is a significant change in each of the four distributions.

3.3 Results and Evaluations

3.3.1 The Change of Pillai Scores

In order to make a better comparison, Pillai scores between these four following pairs are compared:

1. IH + a nasal consonant vs. IH + a non-nasal consonant. (represented as PIN-PIT)
2. EH + a nasal consonant vs. EH + a non-nasal consonant. (represented as PEN-PET)
3. IH + a nasal consonant vs. EH + a nasal consonant. (represented as PIN-PEN)
4. IH + a non-nasal consonant vs. EH + a non-nasal consonant. (represented as PIT-PET)

Each of the four pairs of clusters generates a Pillai score. After performing forced alignment the second time (Step 2 in Figure 3.1) and re-assigning vowel labels for vowel formant data (Step 4 in Figure 3.1), we would expect some changes in the first three categories above. The goal is to check which ones have changed significantly in these four categories. It is expected that the third category (PIN-PEN) has the most substantial increase, and there should not be a change in the last category (PIT-PET) because in the experiments we did not modify the pronunciations of PIT and PET words.

Table 3.1: Numbers of IH and EH Tokens in All Manners after Modification and Outlier Removal. Only vowel tokens with primary stress are included.

	Total IH Tokens	Total IH Tokens After Modification	Total EH Tokens	Total EH Tokens After Modification
Before Outlier Removal	111738	115518	131944	128081
After Outlier Removal	106121	110169	125317	121193

Table 3.2: Numbers of pin and pen Tokens after Modification and Outlier Removal. Only vowel tokens with primary stress are included.

	Total pin Tokens	Total pin Tokens After Modification	Total pen Tokens	Total pen Tokens After Modification
Before Outlier Removal	28313	32137	28719	24853
After Outlier Removal	26016	30100	26693	22571

In order to analyze the results, some filters were firstly applied to the two sets of results on formant extraction (Value Set 1 and Value Set 2 in Figure 3.1). Only vowel tokens with the primary stress markers were kept.[12] We also normalized the F1 and F2 values so these values were no longer sensitive to the sex of the speaker. Usually, the F1 and F2 values of male speakers are lower than those of female speakers, so in order to remove this effect and conduct better comparisons, we applied the z-score normalization (also called Labanov normalization in phonetics) on the formant values.

Next, we removed the outliers for each vowel based on the Mahalanobis Distance([37]). If a token's Mahalanobis Distance was outside the 95% quantile of the distribution, then we considered it as an outlier. The total numbers of IH and EH tokens (in all manners and in the nasal manner) after modification and outlier removal can be found in Table 3.1 and Table 3.2.

After normalization and outlier removal, four values that represent the changes of the Pillai scores between four different pairs were calculated for each speaker. Out of the 64 speakers, 62 speakers show an increase in the Pillai score between pin and pen. The only speakers (657X and 312) who show a decrease are also among the speakers with the high

[12] The main difference between stressed and unstressed vowels is that the stressed ones are longer and more carefully articulated, whereas the unstressed ones usually show reduction. Hence we only kept the tokens with primary stress markers in the data.

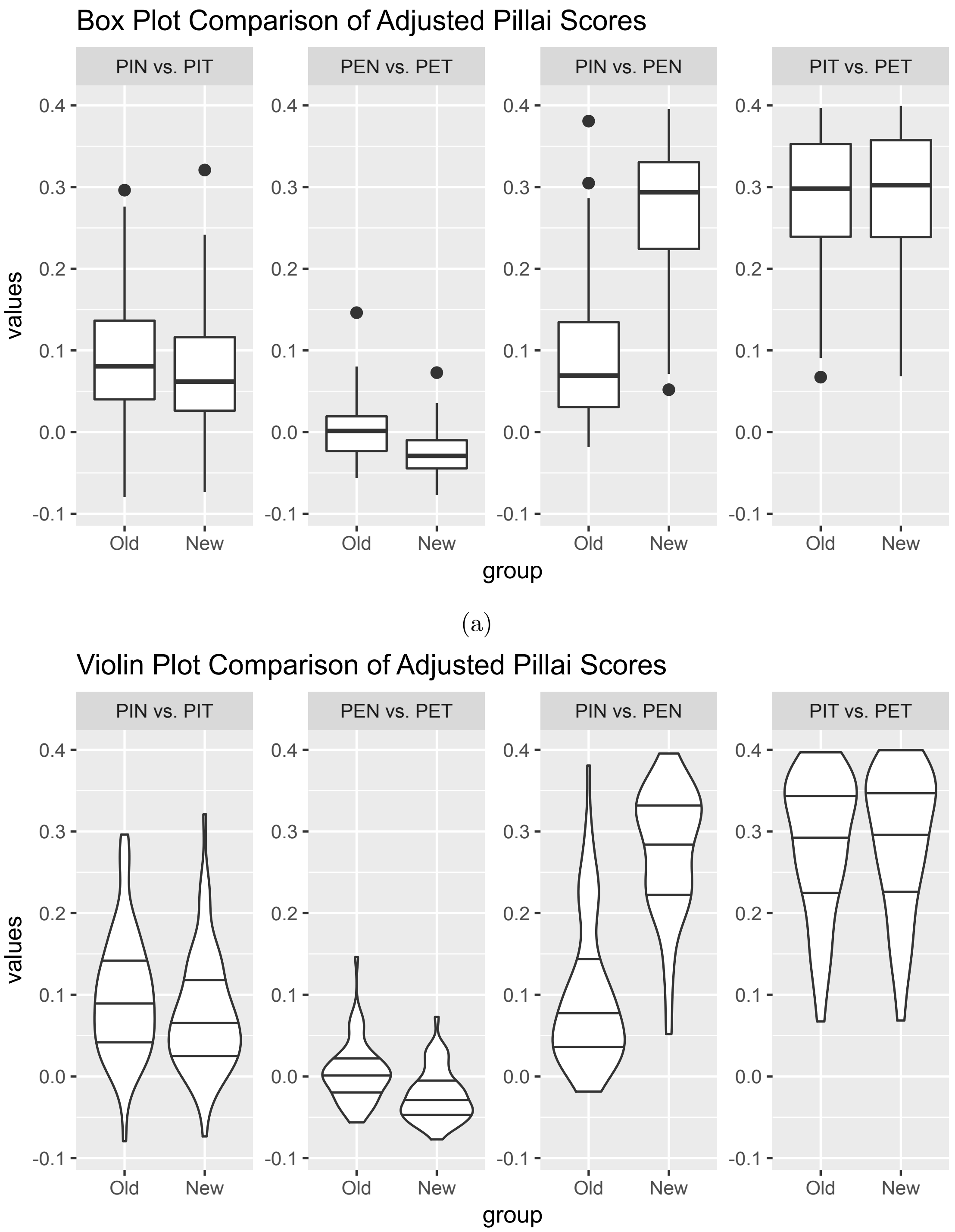
PIN-PEN Pillai score originally.[13] Detailed numerical comparisons can be found in Table B.1. Figures 3.3 represents another way of visualizing the changes of the Pillai scores. Each of figures in 3.3 contains four subplots, which represent four Pillai scores before and after the modification (represented as *old* and *new* columns). The horizontal line in the box plot shows the mean value of the Pillai scores of all selected speakers. In the violin subplots, the horizontal lines in each column represent the value at the 25, 50, and 75 quantiles.

Results show that the Pillai scores change significantly in two out of the four phonetics contexts — PEN-PET and PIN-PEN. The p-values of both categories are less than 0.0001, which show that the Pillai Scores of DASS speakers in these two categories change significantly. This is consistent with our hypotheses on the changes of formant values after modifying the pronunciation dictionary based on the prenasal merger.

3.3.2 Changes of Formant Values at the Individual Level

However, the Pillai score is only a summary statistic and cannot reveal how much each speaker's F1 and F2 values change in different phonetic contexts. Therefore, we calculated the relative difference and conduct t-tests on F1 and F2 values of PIN, PIT, PET, and PEN categories for each speaker in order to check whether there would be a significant change in F1 or F2 values of these categories. The difference after modifying the pronunciation dictionary can be seen in Figure 3.4. Two categories that change the most are the F1 and F2 values for PEN tokens. The F1 values of PEN tend to increase, which means PEN tokens are on average lower in the vowel space. In contrast, the F2 values of PEN tend to decrease, which means PEN tokens are more towards the back of the vowel space. Both F1 and F2 values in the second run of forced alignment are more in conformity with the standard American Vowel Space illustrated in Figure 1.3. Besides the F1 and F2 values of PEN, the F1 and F2

[13] The speaker with the highest score originally is Speaker 657X and the decrease of his PIN-PEN Pillai score (-0.07) is also the greatest. 657X is European American male from Louisiana and born in 1897. He learned French before learning English.



(a)

(b)

Figure 3.3: Box and Violin Plots of Side-by-side Comparisons of Pillai Scores Before and After Forced Alignment.

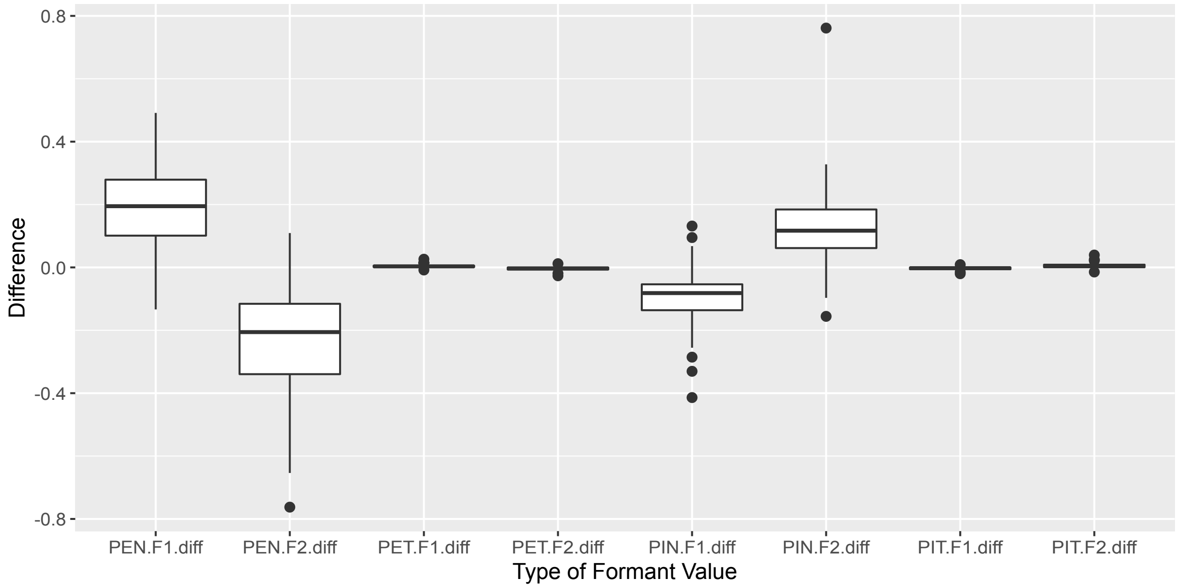


Figure 3.4: Changes of F1 and F2 values in Different Contexts.

values of PIN tokens have also changed, although the change is not as significant as that of PEN. For PIN tokens, the average F1 values decrease whereas the average F2 values increase. All other categories, including F1 and F2 values for PIT and PET tokens, do not have any significant change because we did not make any changes in the pronunciation dictionary for these words.

In addition, t-tests for each speaker can reveal how many speakers have shown a significant increase in each of the eight categories. Consistent with Figure 3.4, none of the 64 speakers show a significant change in the F1 or F2 values of PIT or PET. 48 and 45 speakers have a significant increase in the F1 and F2 values of PEN, respectively. 24 speaker have shown a significant change in the F1 values of PIN, and 26 speakers have shown a significant change in the F2 values of PIN. Detailed results across 64 speakers can be found in Table B.2. Combining the results shown in Figure 3.4 and t-tests, we can conclude that modifying the pronunciation dictionary changes the formant values of PEN the most, and at the same time,

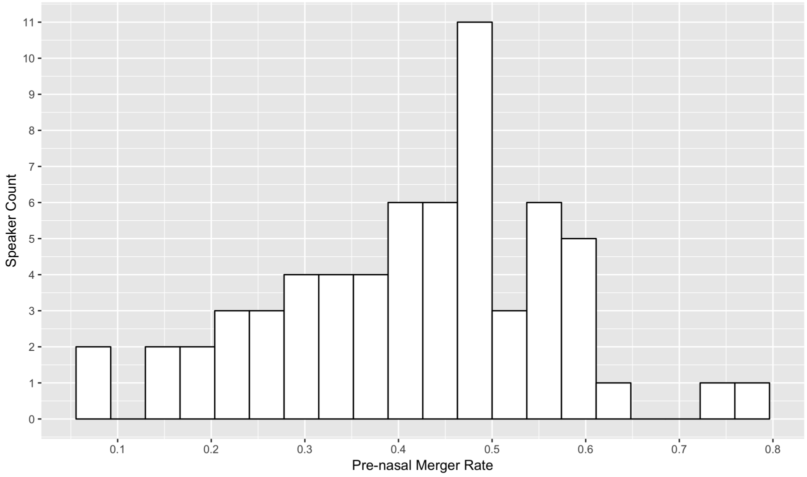


Figure 3.5: Distribution of PEN-PIN RR among 64 Speakers.

the formant values of PIN remain more stable. All these results are also consistent with our expectation on what would happen if we modify the pronunciation dictionary based on the prenasal merger.

3.3.3 Reclassification Rates

Proportions of merged PEN tokens can provide valuable information on the prenasal merger. Allegedly, a high proportion of changed instances (PEN to PIN) indicates a high degree of the prenasal merger for a speaker. Figure 3.5 shows the distribution of the proportions of PEN being re-classified as PIN, which we call PEN-to-PIN Reclassification Rate (PEN-PIN RR).

Figure 3.6 presents scatterplots showing how the PEN-PIN Reclassification Rates correlate with the original PIN-PEN Pillai Score. We use the Pearson Correlation Coefficient to provide quantitative measures of the correlation. The results show that there is a moderate correlation (R = -0.506, $p < 0.0001$) between PEN-PIN RR and the original PIN-PEN Pillai Score, which means a speaker with a higher PEN-PIN RR tend to have a low origi-

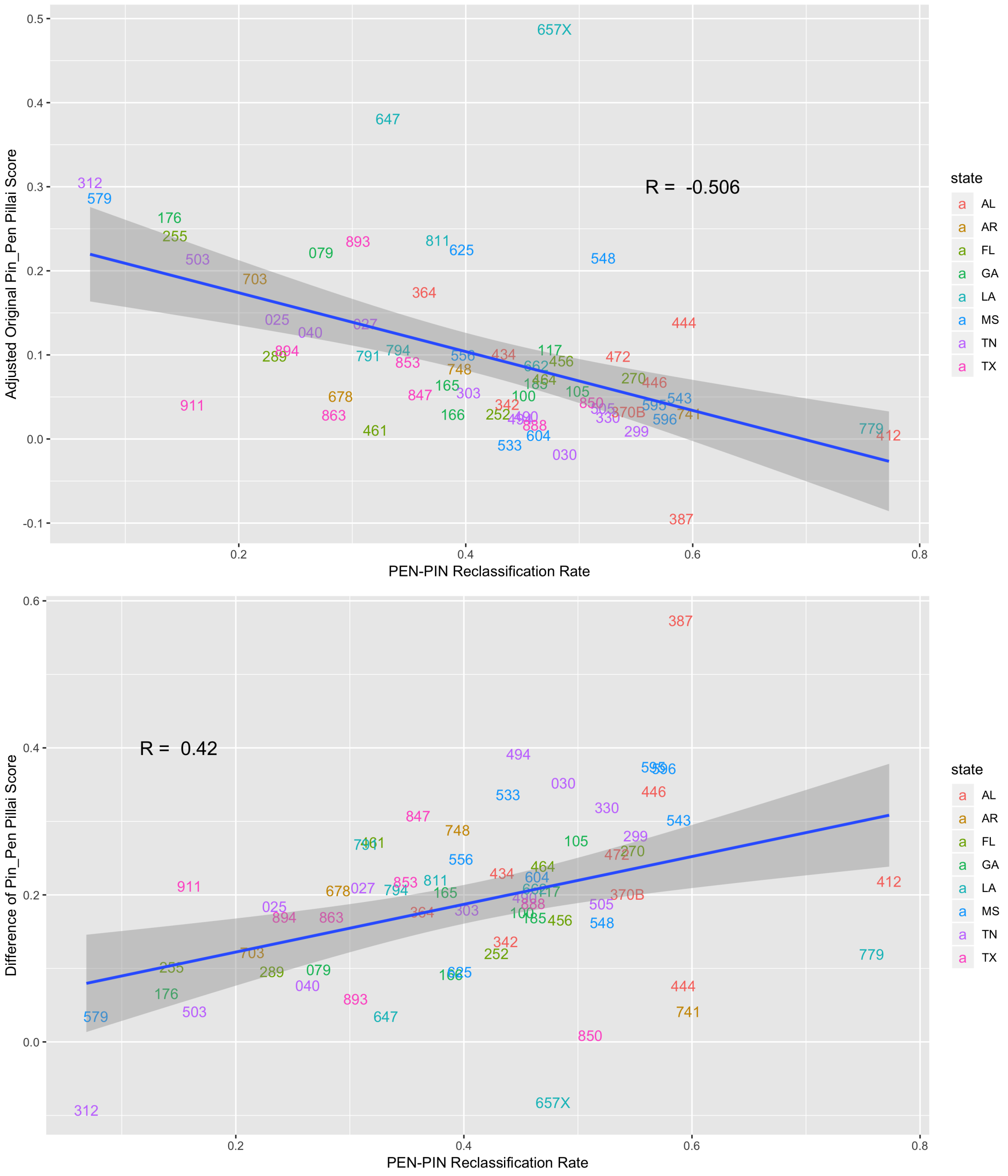


Figure 3.6: (a) Scatterplot of PEN-PIN Reclassification Rate and Old PIN-PEN Pillai Score; (b) Scatterplot of PEN-PIN Reclassification Rate and the Difference between New and Old Pillai Scores.

nal PIN-PEN Pillai Score and vice versa. One reason that the correlation is not extremely high is that Pillai scores tend to underestimate small differences than other metrics such as Euclidean distances.[14] Another possible reason is that when classifying each phone, the forced aligner uses MFCC features, whereas in formant extraction, Praat or FAVE uses LPC (linear-predictive-coding) features, which is related to but different than MFCC features.[15] Another reason that might contribute to this discrepancy is the noisy nature of the data. Further research can be done on the difference and relationship between MFCC and LPC features in detecting sociolinguistic variations and how well they perform on clean laboratory speech.

The change of each speaker's PIN-PEN Pillai score after the modification also correlates with each speaker's original PIN-PEN Pillai score (Pearson Correlation Coefficient R = -0.68). Figure 3.7 is a scatterplot showing the correlation. Speakers who show high degrees of merger (i.e. with low original scores) tend to have a greater increase in their Pillai scores after the modification. In contrast, speakers who originally have high Pillai scores tend to have a smaller increase (or even a decrease).

As an important part of our results, the PIN-PEN RR of each speaker, which indicates the proportion of PIN tokens being reclassified to PEN, is also examined. We studied whether and how it tracked any other statistics of formant values. For a majority of the speakers, the PEN-PIN RR is higher than the PIN-PEN RR of the speaker. This is represented in Figure 3.8. There were 15 speakers whose PIN-PEN RRs were higher than their PEN-PIN RRs. One factor that contributes to this result could be the noise in the audio processing and formant extraction. Besides the noise, the factor that likely contributes the most to the abnormal outcome was that the average F1 and F2 values of originally classified PIN tokens of these speakers diverged from the majority, which means that the acoustic space for each of these

[14] For detailed discussions on these disadvantages of Pillai Score, See [6], [9], and [38].

[15] For estimating formant values through LPC features, see [39].

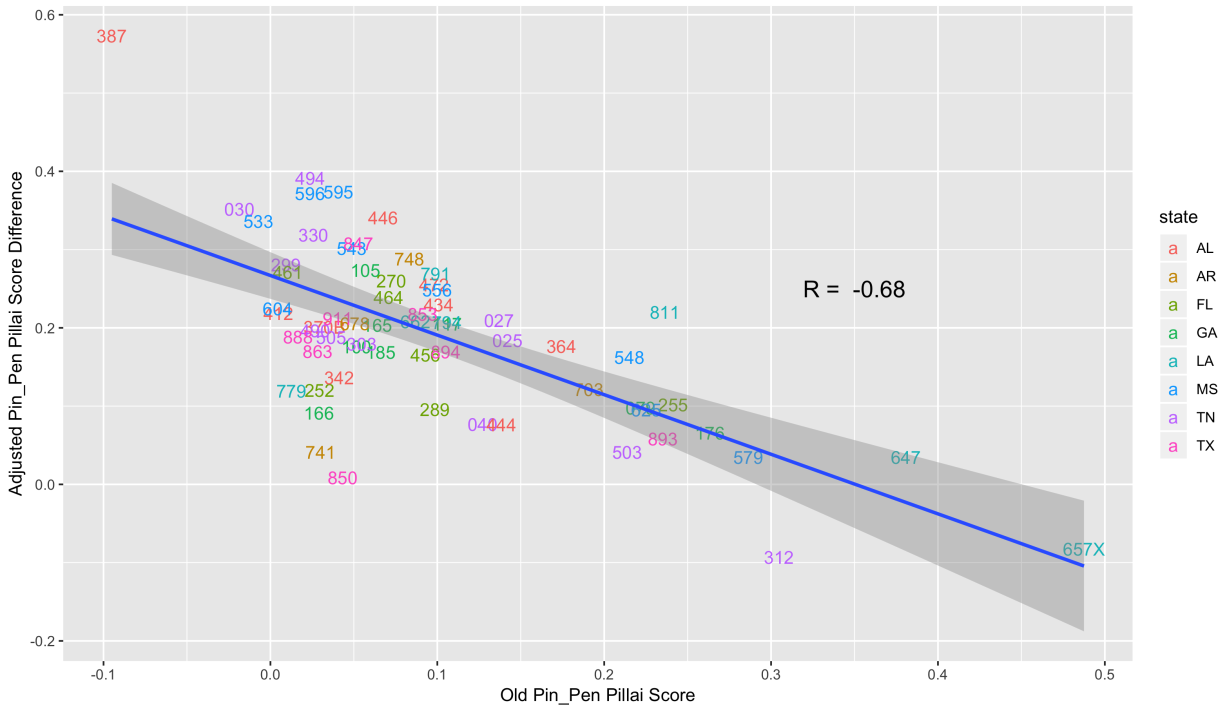


Figure 3.7: Correlation between a Speaker's Adjusted Pillai Score before Modifying the Pronunciation Dictionary and the Change after Modifying the Pronunciation Dictionary. This chart shows that speakers with low original Pillai scores tend to have a higher increase in the results from the modified forced alignment.

outlier speakers was likely to be different from the standard one. Figure 3.9 and Figure 3.10 show how a speaker's PEN-PIN RR and PIN-PEN RR are related to his or her average F1 and F2 values of PEN and PIN tokens in the vowel space, respectively. We find that speakers whose average F1 and F2 values of PEN or PIN tokens deviate from the majority (i.e. they are outliers) tend to have a high PEN-PIN RR or PIN-PEN RR, respectively. The reason is that, when we used a pre-trained acoustic model trained on the standard American English[16] in forced alignment, there were likely to be more inaccuracies among the vowel tokens of these speakers because the model is trained on clean data without many outliers in terms of speech variations. Specifically, we used Pearson Correlation Coefficient to examine the correlations, and we found that a high PIN-PEN RR is correlated with a high F1 value (R = 0.466, p = 0.0001) and a low F2 value (R = -0.45, p = 0.0002) of PIN tokens. The Pearson Correlation Coefficient and the accompanied p-values between statistics of formant values and Reclassification Rates (including PEN-PIN RR and PIN-PEN RR) are listed in Table 3.3.

In performing these experiments, we also noticed that one of the main advantages of using forced alignment to identify the prenasal merger was its efficiency. Our results show that in sequential processing, modified forced alignment only takes about 10% of the processing time of formant extraction.

3.4 Clustering Analysis of Vowel Formant Values

By examining the statistics such as the Pillai Score and the Euclidean Distance, we have shown that there is a better separation between the clusters of PIN and PEN tokens after forced alignment with a suitably modified pronunciation dictionary. As we have argued, the better separation is an indicator that our modification yields transcriptions that are more accurate because of the increased consistency with the vowel space of the standard American

[16] LibriSpeech corpus. See [32].

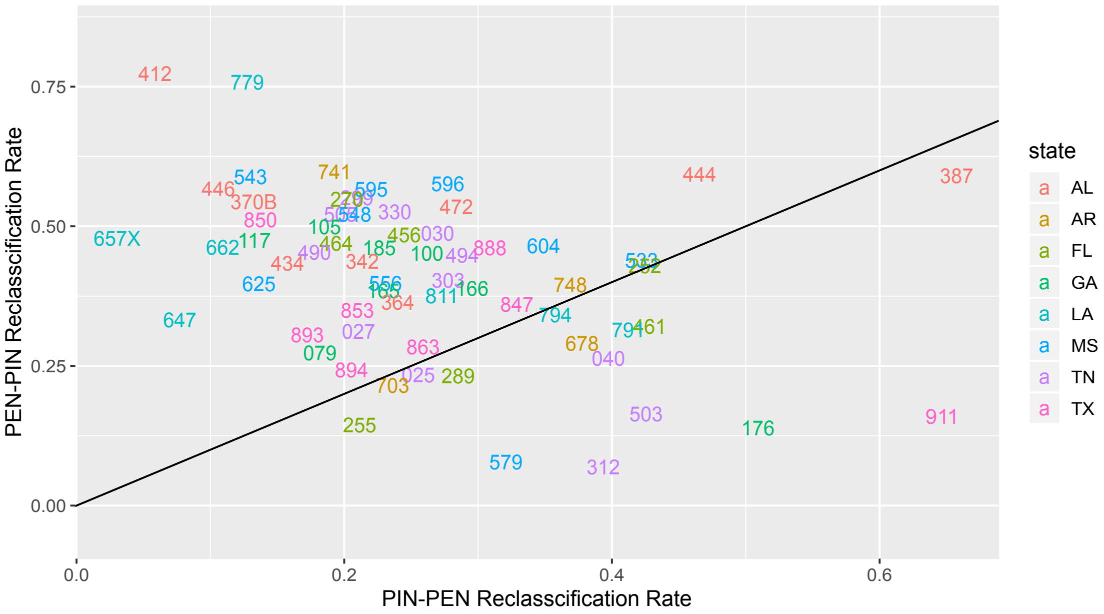


Figure 3.8: PIN-PEN RR and PEN-PIN RR for each DASS Speaker. The Diagonal Line indicates the slope $y = x$. That most points are above the line indicates that most speakers PEN-PIN RRs are higher than their PIN-PEN RRs.

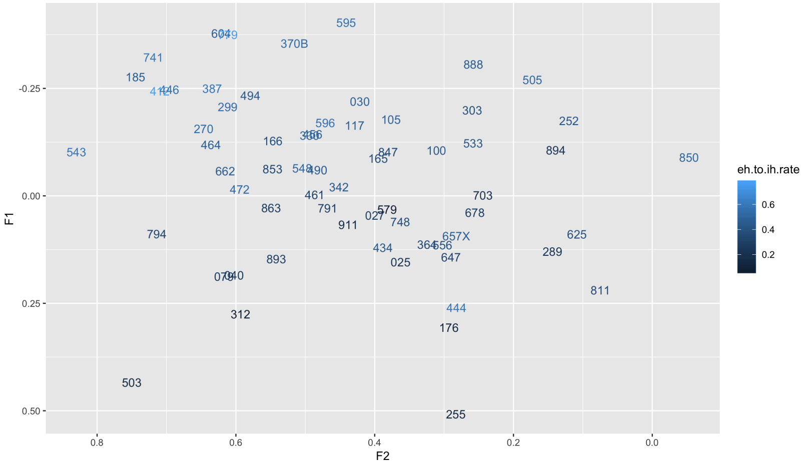


Figure 3.9: PEN-PIN RR and the Average F1 and F2 Values of PEN Tokens for Each Speaker.

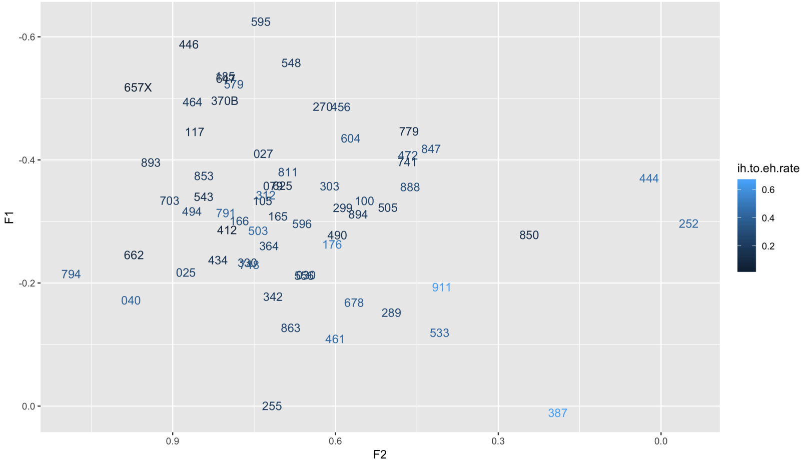


Figure 3.10: PIN-PEN RR and the Average F1 and F2 Values of PIN Tokens for Each Speaker.

Table 3.3: Pearson Correlation Coefficient between Statistics of Formant Values and Reclassification Rates. The numbers in the parentheses indicate the p-values. Bold indicates p <0.05.

	PEN-PIN RR	PIN-PEN RR
PEN-PIN Pillai	**-0.506(<0.0001)**	-0.205(0.104)
PIN-PIT Pillai	-0.02(0.821)	**0.369(0.0027)**
PIT-PET Pillai	-0.108(0.396)	0.006(0.963)
PEN-PET Pillai	**0.534(<0.0001)**	**-0.287(0.022)**
Change PEN-PIN Pillai	**0.42(0.0006)**	0.196(0.1207)
Change PEN-PET Pillai	**-0.61(<0.0001)**	0.09(0.482)
PEN-PIN ED	**-0.506(<0.0001)**	0.196(0.121)
Change PIN-PEN ED	**-0.512(<0.0001)**	0.1523(0.2296)

	PEN-PIN RR	PIN-PEN RR
PEN F1	**-0.71(<0.0001)**	0.0276(0.8289)
PEN F2	**0.25(0.0442)**	0.1273(0.3162)
Change PEN F1	**0.73(<0.0001)**	**-0.2833(0.023)**
Change PEN F2	**-0.74(<0.0001)**	0.1734(0.171)
PIN F1	**-0.28(0.024)**	**0.466(0.0001)**
PIN F2	-0.20(0.109)	**-0.45(0.0002)**
Change PIN F1	**-0.28(0.024)**	**0.466(0.0001)**
Change PIN F2	-0.20(0.109)	**-0.45(0.0002)**

English. In this section, we provide more evidence for the claim that our modification yields better phonetic transcriptions of PIN and PEN tokens.

If manually-transcribed vowel labels could be definitively attached to each phone token of an audio corpus, then we can use these labels as the ground truth to evaluate the accuracy of the phone transcriptions, but in reality, there are no such ground-truth labels for DASS. Thus, we use K-means clustering to approximate the ground-truth lables of IH and EH tokens and measure whether there are any improvements in precision, recall, and the harmonic mean of precision and recall[17] after the modification.

K-means clustering has been used in identifying outliers in the formant data and evaluating speakers' judgments on their own speech production ([40]). In our work, in order to examine whether the modified forced alignment provides better phonetic transcriptions of PIN and PEN tokens, we perform K-means Clustering (K = 2) on all IH and EH tokens (including PIN, PEN, PIT, and PET tokens), which gives us two clusters in the vowel formant space – one higher and fronter (i.e., low F1, high F2 values), and the other lower and backer (i.e., high F1, low F2 values). We evaluate whether the modified forced alignment generates better phonetic transcriptions by checking whether these transcriptions (i.e., Result 2 in Figure 3.1) correspond more closely to the clusters generated by K-means clustering.[18]

One thing to note is that K-means clustering, as an unsupervised clustering technique, by itself does not assign labels for each cluster. We assign labels to the results of K-means clustering in a post-hoc fashion. Based on the common notion of the vowel space of American English, the *higher* cluster (i.e., the cluster with a lower mean F1 value) is assigned the label IH and the *lower* cluster is assigned the label EH.

Detailed plots of each speaker's K-means clustering result are in Appendix D. In order to conduct a better comparison, we remove the results of 8 DASS speakers based on post-hoc

[17] This is also called the F1 score, which is different from the F1 formant value.

[18] During the process of this project, we also tried another clustering method – Gaussian Mixture Clustering. And the results are largely similar with the K-means clustering results provided here.

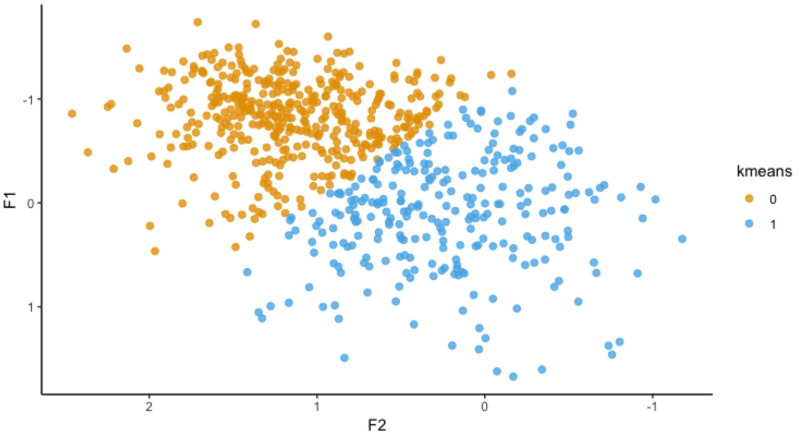
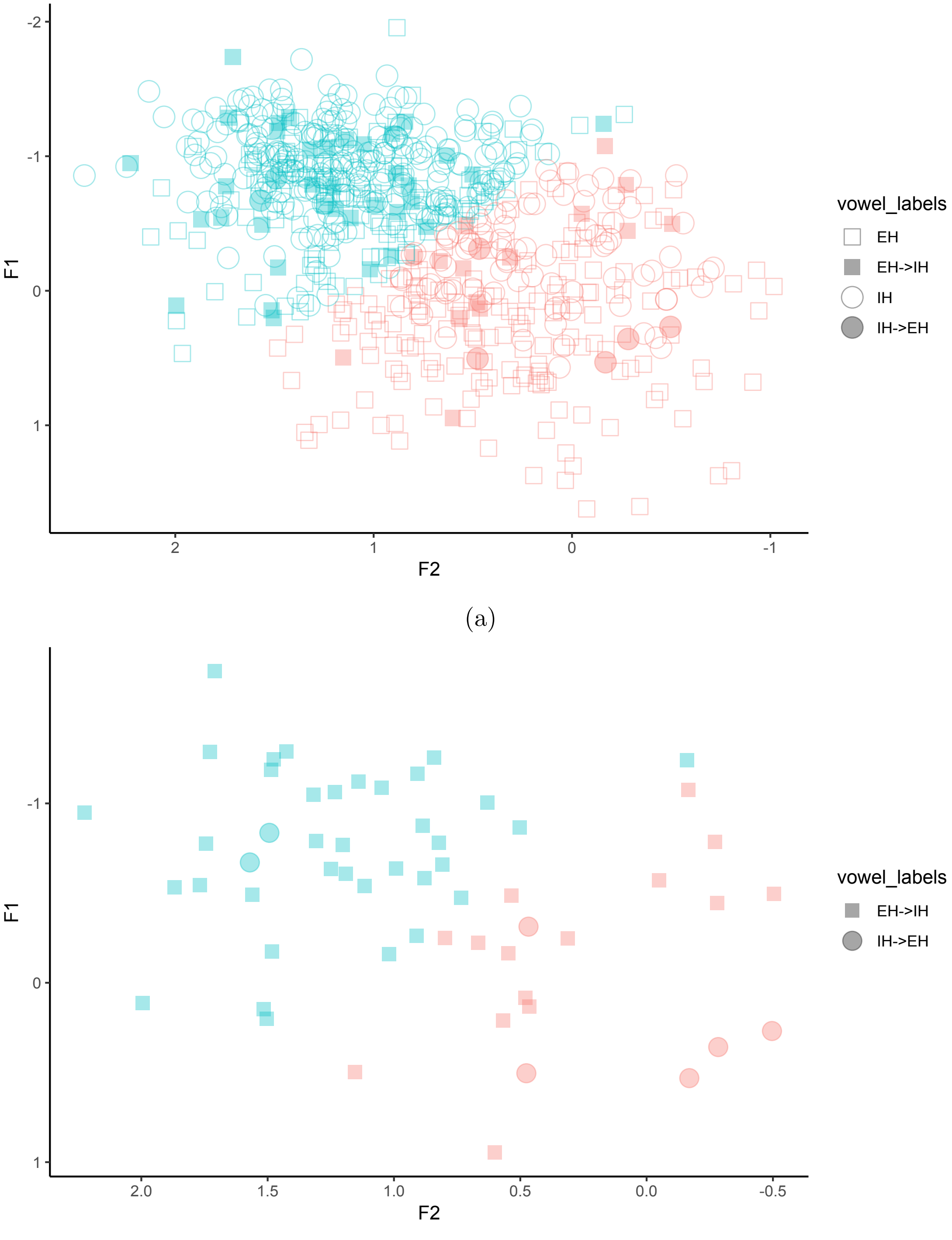


Figure 3.11: K-means Clustering Results of PIN and PEN Formant Values on DASS Speaker 446 (European American Female from Alabama; Born in 1917).

visual inspections. These speakers are 252, 364, 444, 456, 472, 548, 779, and 911.[19] We remove 548 because in her results there are missing formant values in a certain range of frequencies. In the K-means clustering results of the other 7 speakers, we notice that one cluster has *high* F1 and *high* F2 values and the other has *low* F1 and *low* F2 values, which is inconsistent with the patterns specified by the common understanding of vowel space in Figure 1.3. Therefore, we consider the results from these 8 speakers to be skewed and remove them in calculating the accuracy metrics.

Based on the results of the remaining 56 speakers, Table 3.4 shows the metrics before and after modifying the pronunciation dictionary in terms of classifying all PIN and PEN tokens. Specifically, the numbers in the table represents the mean value of the weighted averages of precision, recall, and F1 score for the 56 selected speakers. There is a substantial increase in all these three categories, which indicates that after the modification, PIN and PEN tokens are classified in better correspondence with the K-means clustering results as well as the common understanding of the American English vowel space.

[19] Detailed demographics information of these speakers can be found in [21].



(a)

(b)

Figure 3.12: K-means Clustering Results of DASS Speaker 446. X axis and Y axis are normalized F2 and F1 scores, respectively. Blue and red data points refer to the labels of IH and EH from 2-means clustering.

Table 3.4: Weighted averages of precision, recall, and F1 score in classifying PIN and PEN tokens before and after modifying the pronunciation dictionary, averaged over the 56 selected DASS speakers.

	Precision	Recall	F1 Score
Before	0.5812	0.5662	0.5665
After	0.6993	0.6683	0.6627

To better illustrate the locations of reclassified tokens, in the rest of section we show the results of DASS Speaker 446.[20] Figure 3.12 illustrates the K-means clustering results of DASS Speaker 446. Both plots are from the same set of results, and in order to better illustrate the change, we select all tokens with updated labels and plot Figure 3.12 (b), which is a subset of Figure 3.12 (a). Different colors represent labels from K-means clustering, and different shapes represent labels from two runs of forced alignment. For all the dark squares representing vowel measurements whose labels have been changed from EH to IH, the majority of them (36 instances) are in the blue cluster (labeled IH according to 2-means clustering), and only 15 instances of the dark squares are in the red cluster (labeled EH according to 2-means clustering). These results show that forced alignment with the modified pronunciation dictionary re-assigns the IH label to vowel tokens that are higher in the vowel space, which further explains the increase in Table 3.4.

[20] European American Female from Alabama; Born in 1917.

Chapter 4

Comparisons on the KL-Divergence between Different Acoustic Models

In this chapter, we provide a more detailed investigation on GMM, HMM, and how they can be used in combination with KL-Divergence to reveal the detailed differences between the prenasal allophones of IH and EH. These two allophones correspond to PIN and PEN tokens recognized by the forced aligner. Our results are also helpful in understanding how the ASR technique is implemented in distinguishing between PIN and PEN tokens.

4.1 Gaussian Mixture Models and Probabilistic Hidden Markov Models

As we have introduced in Chapter 2, Gaussian Mixture Models can be used in acoustic modeling. Each of GMMs used in an ASR system is a mixture of multivariate Gaussian distributions. The formulation of a multivariate distribution of an n-dimensional random vector x is as follows:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{(D/2)}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)] \tag{4.1}$$

The value μ is an n-dimensional mean vector, and Σ is the covariance matrix with size $n \times n$. The values of μ and Σ determine a multivariate Gaussian distribution. This probability density function can be represented as $p(x|\mu, \Sigma)$.

Consequently, a mixture of k Gaussian models can be represented as

$$p_M(x) = \sum_{i=1}^{n} \alpha_i \cdot p(x|\mu_i, \Sigma_i) \tag{4.2}$$

In Equation 4.2, μ_i and Σ_i represent the parameters for the i-th Gaussian component, and another constraint is that for each $i \leq k$, the mixture coefficient $\alpha_i > 0$, and $\sum_{i=1}^{n} \alpha_i = 1$.

In the context of ASR, the dimension of each Gaussian component is 39, which is the same with the dimension of MFCC features. According to Equation 4.1, the covariance matrix Σ for each Gaussian component has a dimension of 39 × 39, which represents the variance between each pair of feature dimensions. But in the practice of speech recognition, if each 39-dimensional Gaussian component has a covariance matrix of size 39 × 39, the mixture model will contain a large amount of parameters (39 × 39 × the number of Gaussian components), which makes the training difficult. A common way to simplify the model and make the model more efficient to train is to assume that the 39 features in each dimension do not covary. Under this assumption, the covariance matrix can be represented by a diagonal matrix.[1] To train a GMM is to find the best parameters μ_i, α_i, and Σ_i in Formula 4.1. Expectation-Maximization (EM) Algorithm ([42]) is often used in the process.

[1] This also corresponds to the fact that MFCC features are uncorrelated. There are also other variations of this step which do not a diagonal covariance matrix. For example, [41] proposes a subspace Gaussian Mixture Model that uses full covariance instead and generates then-state-of-art results. In general, a GMM with a full covariance matrix represents a better acoustic model than one with a diagonal covariance matrix. But mainly for the sake of performance, diagonal covariance matrices are usually used instead.

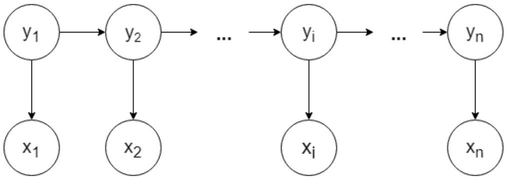


Figure 4.1: Structure of HMM.

In a Hidden Markov Model (HMM), there are two sets of variables: hidden variables and observations. We use $y_1, y_2, ..., y_n$ to denote the hidden variable in a series of n states, and use $x_1, x_2, ..., x_n$ as the observation in each of the n states. The structure of a HMM can be shown in Figure 4.1. There are two important assumptions of an HMM. The first is that the probability of a particular state y_i is only dependent upon the previous state. This is the Markov Assumption, which can be formulated as 4.3 in probabilistic form.

$$P(y_i|y_1...y_n) = P(y_i|y_{i-1}) \tag{4.3}$$

The second assumption is that, the output observation at a particular state x_i is only dependent upon the state that produces the observation y_i, which means different observations x_i are independent to each other. This is the Output Independence Assumption (Equation 4.4).

$$P(x_i|y_1...y_i, x_1, ..., x_{i-1}) = P(x_i|y_i) \tag{4.4}$$

In addition to the two assumptions above, three sets of parameters are also needed for an HMM. First is a transition probability matrix $\mathbf{A} = [a_{ij}]_{N\times N}$, in which

$$a_{ij} = P(y_{t+1} = s_j|y_t = s_i), 1 \leq i, j \leq N$$

N is a discrete variable and it denotes possible states of each y_i. a_{ij} represents the probability of moving from state i at time t to state j at time $t+1$.

The second parameter needed in an HMM is an initial probability distribution $\pi = (\pi_1, \pi_2, ..., \pi_N)$, where

$$\pi_i = P(y_1 = s_i)$$

In the above equation, π_i denotes the probability that the initial state y_1 is s_i. The sum of all π_i is 1.

Another parameter in an HMM is Observation Likelihoods, which are also called emission probabilities. Suppose the observations are discrete, and there are M possible observations $o_1, o_2, ..., o_M$. The Observation Likelihoods can be represented by a matrix $\mathbf{B} = [b_{ij}]_{N \times M}$, in which

$$b_{ij} = P(x_t = o_j | y_t = s_i), 1 \leq i \leq N, 1 \leq j \leq M$$

b_{ij} represent at time t, the probability of the observation o_j given the state s_i. If the observations o are not discrete, then a Probability Density Function can be used to characterize the distribution of o_j, making the HMM a probabilistic model. In the specific context of ASR, a GMM for each observation o_t can be used. For an m-component GMM, the observation variable $\mathbf{x}$ can be calculated as

$$b_i(\mathbf{x}) = p(\mathbf{x}|S = i) = \sum_{m=1}^{M} c_{im} \mathcal{N}(x; \mu_{im}, \Sigma_{im})$$

The topology of a probabilistic HMM can be shown as Figure 4.2. Using the terminology of GMM-HMM models, the process of forced alignment is to find the best sequence of states $S_1, S_2, ..., S_n$ that generates the corresponding sequence of observations $X_1, X_2, ..., X_n$, where S_i and X_i stand for the phone label and the MFCC feature respectively at each time window

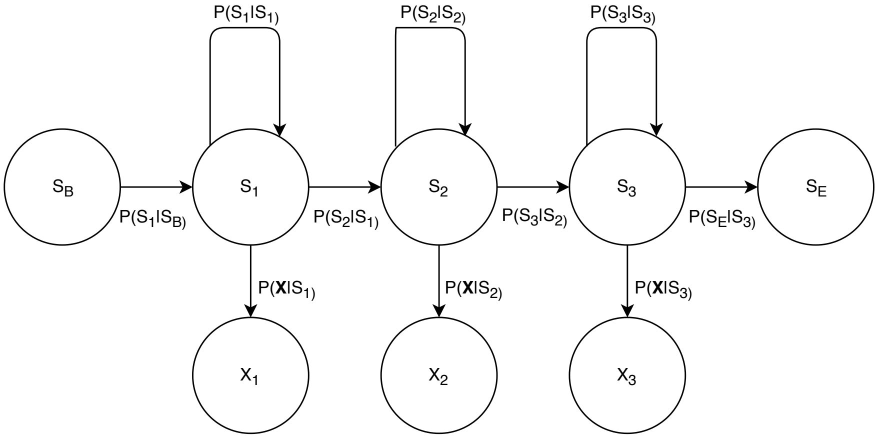


Figure 4.2: Probabilistic HMM.

i.

4.2 KL-Divergence

Widely used for evaluating the difference between two probabilistic distributions, Kullback-Leibler divergence (KL-divergence) is also called relative entropy or information divergence. Given two probabilistic distribution P and Q, the KL-divergence between P and Q is

$$KL(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx \tag{4.5}$$

where $p(x)$ and $q(x)$ are the probabilistic density functions of P and Q. When P and Q are the same distribution, $KL(P||Q) = 0$.

Because KL-divergence can be used to measure the difference between two probabilistic distributions and GMM-HMM acoustic models are probabilistic models, KL-divergence can be used to measure the difference between two different acoustic models, which can be further applied into tasks like speaker identification and verification.[2] Similar techniques could also be applied on feature discoveries of specific phonemes. In order to study g-dropping in American English, [14] trained customized models for two new phonemes, IHN and IHNG, which are through combining the IH vowel with following consonants N and NG respectively and used in identifying whether the g-dropping is present. The authors evaluate the KL-divergence between trained acoustic models of IHN and IHNG from forced alignment and show where these two acoustic models differ the most.

In this section, using a similar method, we apply the Monte-Carlo estimation method in measuring the KL-Divergence between different acoustic models from the two different runs of forced alignment, and examine how the results of KL-Divergence is coherent with the assumption of prenasal merger. Specifically, we can use KL-divergence to measure how the acoustic models of prenasal IH and EH differ in each of the three HMM states. If the KL-divergence between these two models are high, then it means the acoustic models have a high degree of dissimilarity, and vice versa.

One problem for measuring the KL-divergence between different GMMs is that the KL-divergence between two different GMMs is not analytically tractable. In response, different methods are proposed to estimate the KL-divergence between different GMMs. The Monte-Carlo estimation is a common method. Although it is stochastic and has a high computational cost (see [44]), it is widely adopted and its results are also used in evaluating the performance of other approximation methods.[3] [46] performs numerical comparisons the different ways of approximating the KL-divergence between different GMMs and points out the problem of the stochastic nature and slow computational time of the Monte-Carlo estimation. In our work, only a small number of GMMs are compared so the computational

[2] Previous work has trained speaker-specific or speaker-adapted GMM models and used the KL-divergence between these models to evaluate the difference between different speakers, which can be further applied into the tasks of speaker identification and verification. For instance, see [43].

[3] Detailed results can be found in [45].

cost is not an issue. To remedy the stochastic nature of the Monte-Carlo estimation, we fix the random seed to ensure the reproducibility, and for each model comparison, we perform 10^6 iterations to get a more accurate estimation.[4]

4.3 Experiment and Results

In order to compare how the KL-divergence between different acoustic models changes after modifying the pronunciation dictionary based on the prenasal merger, we make additional changes on the pronunciation dictionary. Instead of only adding new pronunciations for words based on the prenasal merger, we keep the same number of entries but add new *phonemes* to capture the variations of vowels EH and IH in different phonetic contexts. Similar methods are used in [16] and [14].[5] In the pronunciation dictionaries used in this experiment, we replace all instances of prenasal EH and IH into EHN and IHN respectively, and keep instances of EH and IH in other phonetic contexts to be the same. By doing so, we are able to get separate GMM models for vowels IH and EH in nasal and non-nasal contexts. Specifically, here is the inequality condition we verify:

$$KL(G_{IHN}||G_{EHN}) < KL(G_{IHNm}||G_{EHNm}) \tag{4.6}$$

G_{IHN}, G_{EHN} refer to the trained acoustic models of prenasal IH and EH *before* modifying the pronunciation dictionary, and G_{IHNm}, G_{EHNm} refer to the trained acoustic models of prenasal IH and EH *after* modifying the pronunciation dictionary. Before the modification,

[4] There are two main reasons why we use 10^6 iterations of approximation. First, such a high number is used in [46] for giving an accurate estimate for KL-divergence so that the estimate can be used as the ground truth for other approximation methods to compare with. Second, we are not comparing a large amount of acoustic GMMs that contain a large amount of components, so the computational cost is not an issue.

[5] In [16], the authors modify the pronunciation dictionary by diving the instances of /l/ into two different phonemes in order to study different realizations of the /l/ in different phonetic contexts. /l/ is divided into two phonemes – L1 and L2, in order to model word-initial /l/ (for example, the word *like*) and word-final /l/ (for example, the word *full*)

the merger is not explicitly considered in the forced alignment process. Consequently, vowel EHN is merged to IHN, which means G_{EHN} are more similar to G_{IHN} (i.e. the KL-divergence is smaller) than the scenario where the merger is explicitly considered.[6] Because in the default implementation of Kaldi and Montreal Forced Aligner, each HMM has three states, by comparing the difference between the changes of KL-divergence in these states before and after modifying the pronunciation dictionary, we can reveal which state (i.e. which part of the vowel duration) endures the most significant amount of change.

Another clarification we need to make here is on the *direction* of KL-divergence. KL-divergence is not symmetric, so in most scenarios, $KL(f||g) \neq KL(g||f)$ and they represent different directions of KL-divergence. In our work, we always examine $KL(G_{IHN}||G_{EHN})$ because we take G_{IHN} to be the true distribution we want G_{EHN} to model. The prenasal merger suggests that many tokens of EHN is merged into IHN in natural speech, so a trained acoustic model of EHN without explicitly considering this possibility would be similar to the model of IHN. $KL(G_{IHN}||G_{EHN})$, which can 4.6 can be expanded as

$$KL(G_{IHN}||G_{EHN}) = \int_{-\infty}^{\infty} p_{G_{IHN}}(x) log \frac{p_{G_{IHN}}(x)}{p_{G_{EHN}}(x)} dx \tag{4.7}$$

In the above equation, $KL(G_{IHN}||G_{EHN})$ is proportionate to $\frac{p_{G_{IHN}}(x)}{p_{G_{EHN}}(x)}$, which means that $KL(G_{IHN}||G_{EHN})$ would be high if $p_{G_{IHN}}(x)$ is high and $p_{G_{EHN}}(x)$ is low. In other words, when $KL(G_{IHN}||G_{EHN})$ is high, it means G_{EHN} does not cover much of the probabilistic space of G_{IHN}. The higher $KL(G_{IHN}||G_{EHN})$ is, the more difference there is between the

[6] There are two different approaches of training. One is to train customized acoustic models of GMMs for each speaker, and the other is to train all speakers together get one acoustic model of GMMs for all of them. The first approach is ideal for identifying the variance among different speakers, but is also more sensitive to the noise in the data. Moreover, for some speakers with less audio data, the results won't be meaningful because there are not enough vowel tokens for training. On the contrary, the second approach delivers more stable results but is not able to reveal differences among individuals. DASS is a noisy dataset and the amount of audios for each speaker highly varies, which, combined with the reasons we have mentioned above, pose unique challenges in building individual acoustic GMM models. Hence in our work, we adopt the second approach.

Table 4.1: KL-Divergence between the GMM Acoustic Models of IHN and EHN.

	HMM State 1	HMM State 2	HMM State 3
Before	17.003	14.154	23.98
After	18.313	175.444	222.766

variance of G_{IHN} and the *variance* of G_{EHN}, which is what we need to examine and compare in studying the prenasal merger.[7]

The results can be illustrated in Table 4.1. We notice that HMM State 3, which is the HMM state closest to the vowel's following consonant, undergoes the most significant change. On the contrary, the KL-divergence of the first HMM state does not change much, which is coherent with our expectation because the dictionary modification is only based on the consonant after the vowel. So we can infer that the acoustic models for the prenasal allophones of IH and EH are mostly similar in State 1 and these models distinguish between the two phones based on the MFCC features in the last two states.

[7] So if we were to optimize G_{EHN} based on Equation 4.7, G_{EHN} would usually be *spanned* in order to cover much of the probabilistic space of G_{IHN}. This is also why forward KL-divergence is also called *mass-covering* KL. On the contrary, if we examine and optimize the acoustic model G_{EHN} based on $KL(G_{EHN}||G_{IHN})$ instead, we would not put emphasis on the probabilistic space where $p_{G_{IHN}}(x)$ is high but $p_{G_{EHN}}(x)$ is low, because $\frac{p_{G_{EHN}}(x)}{p_{G_{IHN}}(x)}$ will be down-weighted by the low value of $p_{G_{EHN}}(x)$. For a detailed introduction on the two directions of KL-divergence, see [47].

Chapter 5

Conclusion and Future Work

At the beginning of this thesis, we stated that our general goal was to use forced alignment to provide an analysis of the prenasal merger in Southern American English. Based on the common understanding of the American English vowel space and theoretical hypotheses on how the formant values in different phonetic contexts would change after modifying the pronunciation dictionary in accordance with the prenasal merger, we set up our experiments and obtained results that are consistent with these hypotheses. Specifically, there was a significant increase of separation between the PEN and PIN clusters. For all speakers in DASS, the proportions of PEN-to-PIN reclassification also significantly correlate with the values of the Pillai Score and the Euclidean Distance between these two clusters. Based on these results, we conclude that forced alignment, as a part of ASR techniques, can be used in identifying and analyzing the prenasal merger in Southern American English.

We also performed clustering analysis on vowel formant values. Our results show that forced alignment based on a suitably modified pronunciation dictionary can provide better phonetics transcriptions than its unmodified counterpart. Moreover, we used KL-divergence to measure the difference between the acoustic models of IH and EH in different phonetic contexts. The results show the prenasal allophones of IH and EH differ mostly in the middle and end of their durations.

In conclusion, this work not only exemplifies the methods of using modified forced alignment to identify sociolinguistic variations, but also introduces new approaches of verifying the results of forced alignment without using manually-labeled phone-level transcriptions. These new approaches, including formant analysis and clustering analysis on vowel formant values, provide researchers with new tools for future work. In terms of future research ideas, similar methods of modification and verification could be applied in identifying other sociolinguistic variations in DASS as well as other speech corpora. For instance, researchers can investigate whether the prensasal merger is present among non-Southern speakers in the Buckeye Corpus ([35]), all of whom are from Columbus, OH. The Buckeye Corpus also contains high-quality recordings, so the results should be less noisy and easier to analyze.[1] In addition, researchers can make other modifications in the pronunciation dictionary for the purpose of investigating other kinds of sociolinguistic variations in DASS. For instance, researchers can use forced alignment with a suitably modified pronunciation dictionary to study the prelateral merger of /iy/ and /i/ (as in the word pair *feel* and *fill*) among DASS speakers. Other projects of interest include interpreting acoustic models from deep-neural-network-based ASR systems, which could help further enhance our understanding on the comparison between different phones.

[1] Because manually-labeled phone-level transcriptions are also available for the Buckeye Corpus, we can use these ground-truth labels to verify the accuracy of customized forced alignment and clustering analysis.

Appendix A

ARPABET Tables

The ARPABET is a selection of symbols used within the Advanced Research Projects Agency Speech Understanding Research (ARPA SUR) project ([48]). It is widely used in the speech recognition community. Specifically, it is used in the CMU Pronunciation Dictionary and other popular pronunciation dictionaries for ASR programs. There are two representations of ARPABET symbols – one character and two characters. In this work, the 2-character version is used. Table A.1 shows the correspondence between IPA symbols and ARPABET symbols, including word examples of these symbols.[1]

[1] This particular version of the ARPABET table and examples are retrieved from `http://www.isle.illinois.edu/sst/courses/minicourse/2005/transcriptions.tex`, which is authored by Mark Hasegawa-Johnson (jhasegaw@uiuc.edu) and Sarah Borys (sborys@uiuc.edu).

Table A.1: ARPABET Table.

IPA	ARPABET 1	ARPABET 2	Example
ɑ	a	AA	father
æ	@	AE	bat
ɔ	c	AO	bought
aw	W	AW	bout
aj	Y	AY	bite
e	e	EY	bait
i	i	IY	beat
o	o	OW	boat
oj	O	OY	boy
u	u	UW	school
ə	x	AX	about
X	x	IX	attribute
ɛ	E	EH	bed
ɪ	I	IH	bit
ʊ	U	UH	book
ʌ	A	AH	buck
h	h	HH	hi
ɦ	h	HV	ahead
l	l	L	lead
l̞	L	EL	bottle
r	r	R	roof
r̞	R	ER	bird
r̞	R	AXR	butter
w	w	W	wall
j	y	Y	yacht
m	m	M	mom
m̞	xm	EM	bottom
n	n	N	new
n̞	N	EN	button
ŋ	G	NG	sing
ŋ̞	xG	ENG	tossing
f	f	F	frank
v	v	V	very
θ	T	TH	think
ð	D	DH	that
s	s	S	silly
z	z	Z	zoom
ʃ	S	SH	shelf
ʒ	Z	ZH	azure
p	p	P	pool
b	b	B	bite
t	t	T	tip
d	d	D	dog
tʃ	C	CH	child
dʒ	J	JH	judge
k	k	K	clip
g	g	G	good
ʔ	q	Q	Batman

Appendix B

Pillai Scores and Results of T-tests Before and After Modifying the Pronunciation Dictionary

Table B.1 shows detailed results on the change of each DASS speaker's Pillai scores in the categories of PIN-PEN and PEN-PET before and after our modification on the pronunciation dictionary. Table B.2 shows the results of t-tests on the change of F1, F2 values in different phonetic contexts for DASS Speakers. These results are able to show which speakers endure significant changes in a particular category of formant values. For instance, 48 and 45 speakers had a significant change in the F1 and F2 values of PEN, respectively. 24 and 26 speakers had shown a significant change in the F1 and F2 values of PIN, respectively. None of the speakers showed a significant change in the F1 and F2 values of PIT or PET, which was to be expected because there is no modification on these words in the pronunciation dictionary.

Table B.1: Comparisons of Adjusted Pillai Scores Before and After Modifying the Pronunciation Dictionary in Accordance with the Prenasal Merger.

speaker	PEN-PET	Old PEN-PET	PIN-PEN	Old PIN-PEN	Diff PEN-PET	Diff PIN-PEN
25	-0.0573	-0.0425	0.3263	0.1421	-0.0148	0.1842
27	-0.0974	-0.026	0.3467	0.137	-0.0714	0.2097
30	-0.0444	0.0161	0.3335	-0.0185	-0.0605	0.352
40	-0.0459	-0.0418	0.2037	0.1271	-0.0041	0.0766
79	-0.0572	-0.0262	0.3199	0.2219	-0.031	0.098
100	-0.1023	-0.039	0.2274	0.0516	-0.0633	0.1758
105	-0.0371	0.0138	0.3304	0.0569	-0.0509	0.2735
117	-0.0341	0.026	0.3109	0.1059	-0.0601	0.205
165	-0.1534	-0.0563	0.2675	0.0641	-0.0971	0.2034
166	-0.0578	-0.0259	0.1206	0.0292	-0.0319	0.0914
176	-0.0027	-0.003	0.3293	0.2633	0.0003	0.066
185	0.0141	0.0665	0.235	0.0661	-0.0524	0.1689
252	-0.0596	-0.0412	0.1496	0.0295	-0.0184	0.1201
255	-0.1216	-0.1028	0.3433	0.2414	-0.0188	0.1019
270	-0.0616	0.0571	0.3326	0.0726	-0.1187	0.26
289	-0.0402	-0.0241	0.1942	0.0986	-0.0161	0.0956
299	-0.1053	0.0007	0.2898	0.0094	-0.106	0.2804
303	-0.0241	-0.0016	0.234	0.0548	-0.0225	0.1792
312	-0.0135	-0.0455	0.2118	0.3048	0.032	-0.093
330	-0.0415	0.0181	0.3444	0.0258	-0.0596	0.3186
342	-0.0222	0.0082	0.1775	0.0412	-0.0304	0.1363
364	-0.0405	-0.011	0.3509	0.1743	-0.0295	0.1766
370B	-0.0491	0.0048	0.2335	0.0325	-0.0539	0.201
387	-0.2153	-0.028	0.4783	-0.095	-0.1873	0.5733
412	0.0326	0.1462	0.2232	0.0048	-0.1136	0.2184
434	-0.1137	-0.0363	0.3304	0.1009	-0.0774	0.2295
444	-0.0232	-0.0048	0.2146	0.1384	-0.0184	0.0762
446	-0.0545	0.0034	0.4086	0.0677	-0.0579	0.3409
456	-0.027	0.0024	0.2587	0.093	-0.0294	0.1657
461	-0.1181	-0.0514	0.2817	0.0105	-0.0667	0.2712
464	-0.077	-0.024	0.3099	0.0709	-0.053	0.239
472	-0.019	0.0267	0.3534	0.0982	-0.0457	0.2552
490	-0.1179	-0.0269	0.2232	0.0269	-0.091	0.1963
494	-0.0437	0.0299	0.4153	0.0239	-0.0736	0.3914
503	-0.1007	-0.1269	0.2551	0.2138	0.0262	0.0413
505	-0.0628	-0.0081	0.2243	0.0365	-0.0547	0.1878
533	-0.0608	-0.0083	0.3292	-0.0072	-0.0525	0.3364
543	-0.0444	0.0124	0.3507	0.0489	-0.0568	0.3018
548	-0.016	0.0231	0.3775	0.2151	-0.0391	0.1624
556	-0.0237	0.0078	0.3486	0.0999	-0.0315	0.2487
579	0.0147	0.0023	0.3212	0.2864	0.0124	0.0348
595	-0.0389	0.0221	0.4149	0.0409	-0.061	0.374
596	-0.0114	0.0143	0.3955	0.0238	-0.0257	0.3717
604	0.0331	0.0668	0.2291	0.0045	-0.0337	0.2246
625	0.0728	0.0414	0.3204	0.2253	0.0314	0.0951
647	-0.0311	-0.0089	0.4153	0.3808	-0.0222	0.0345
657X	-0.0504	-0.0115	0.4052	0.4875	-0.0389	-0.0823
662	0.0356	0.0683	0.2952	0.087	-0.0327	0.2082
678	-0.0094	0.0085	0.2563	0.0507	-0.0179	0.2056
703	0.0276	0.0197	0.312	0.1908	0.0079	0.1212
741	-0.0155	0.0057	0.0713	0.0301	-0.0212	0.0412
748	-0.025	0.0026	0.3716	0.0834	-0.0276	0.2882
779	0.0306	0.0803	0.1317	0.0126	-0.0497	0.1191
791	0.0186	0.0575	0.3678	0.099	-0.0389	0.2688
794	-0.0398	-0.0203	0.313	0.1058	-0.0195	0.2072
811	-0.0572	-0.0063	0.4566	0.2363	-0.0509	0.2203
847	-0.0378	-0.0061	0.3603	0.0527	-0.0317	0.3076
850	-0.044	-0.0316	0.0519	0.0433	-0.0124	0.0086
853	0.0072	0.0239	0.3085	0.0912	-0.0167	0.2173
863	-0.0119	0.0122	0.1981	0.0284	-0.0241	0.1697

Table B.2: Results of T-tests on the Change of F1 and F2 Values for Different Speakers. Numbers in the parentheses are the p values. Cells in bold indicate significant changes ($p < 0.05$).

Speaker	NasalEH_F1	NasalEH_F2	NonNasalEH_F1	NonNasalEH_F2	NasalIH_F1	NasalIH_F2	NonNasalIH_F1	NonNasalIH_F2
025	**-4.2345(0.0000)**	**3.0668(0.0022)**	0.0091(0.9927)	-0.0102(0.9919)	**2.5951(0.0096)**	-1.6309(0.1031)	0.0051(0.9959)	-0.0040(0.9968)
027	**-2.9921(0.0029)**	**3.5270(0.0004)**	0.0000(1.0000)	0.0000(1.0000)	0.8008(0.4235)	**-2.2450(0.0251)**	0.0000(1.0000)	0.0000(1.0000)
030	**-7.9814(0.0000)**	**3.4123(0.0007)**	-0.0086(0.9932)	-0.0248(0.9802)	**4.3074(0.0000)**	**-2.2407(0.0254)**	0.0459(0.9634)	0.0100(0.9920)
040	0.6157(0.5389)	0.7786(0.4372)	0.0000(1.0000)	0.0000(1.0000)	-0.4685(0.6402)	-0.9743(0.3316)	0.0000(1.0000)	0.0000(1.0000)
079	-1.8593(0.0633)	**2.2801(0.0228)**	-0.0263(0.9790)	-0.0163(0.9870)	0.7630(0.4456)	-1.8100(0.0705)	0.0004(0.9997)	0.0186(0.9851)
100	-0.6193(0.5366)	**2.3244(0.0213)**	0.0000(1.0000)	0.0000(1.0000)	-0.0557(0.9556)	-1.3215(0.1878)	0.0000(1.0000)	0.0000(1.0000)
105	**-4.7022(0.0000)**	**4.5068(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	1.3867(0.1661)	-1.8659(0.0626)	0.0000(1.0000)	0.0000(1.0000)
117	**-3.9560(0.0001)**	**2.5923(0.0100)**	0.0265(0.9788)	0.0085(0.9932)	0.9079(0.3643)	-0.0797(0.9365)	0.0068(0.9946)	-0.0223(0.9822)
165	**-3.2093(0.0015)**	0.6995(0.4848)	-0.0441(0.9648)	0.0068(0.9946)	1.9238(0.0551)	-0.3111(0.7559)	0.0558(0.9555)	-0.0239(0.9809)
166	**-2.3177(0.0208)**	0.9279(0.3539)	0.0000(1.0000)	0.0000(1.0000)	1.4121(0.1584)	-0.5175(0.6050)	0.0000(1.0000)	0.0000(1.0000)
176	**2.3859(0.0172)**	0.0046(0.9963)	0.0000(1.0000)	0.0000(1.0000)	1.8897(0.0595)	-1.6763(0.0943)	0.0000(1.0000)	0.0000(1.0000)
185	**-4.4713(0.0000)**	**2.4416(0.0149)**	0.0000(1.0000)	0.0000(1.0000)	1.8491(0.0647)	-1.5808(0.1142)	0.0000(1.0000)	0.0000(1.0000)
252	-1.3816(0.1684)	**2.0453(0.0419)**	0.0000(1.0000)	0.0000(1.0000)	1.5518(0.1219)	**-2.0997(0.0367)**	0.0000(1.0000)	0.0000(1.0000)
255	-0.8065(0.4202)	0.4448(0.6565)	0.0000(1.0000)	0.0000(1.0000)	1.5215(0.1286)	-0.8998(0.3686)	0.0000(1.0000)	0.0000(1.0000)
270	**-3.1811(0.0018)**	1.4818(0.1405)	0.0000(1.0000)	0.0000(1.0000)	0.8035(0.4227)	-0.8802(0.3798)	0.0000(1.0000)	0.0000(1.0000)
289	-0.7141(0.4753)	0.9039(0.3662)	0.0000(1.0000)	0.0000(1.0000)	1.3744(0.1695)	-1.7943(0.0730)	0.0000(1.0000)	0.0000(1.0000)
299	**-8.0601(0.0000)**	**8.1019(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	**3.2306(0.0013)**	**-4.7362(0.0000)**	0.0000(1.0000)	0.0000(1.0000)
303	**-4.0275(0.0001)**	**2.5189(0.0119)**	0.0000(1.0000)	0.0000(1.0000)	**2.8400(0.0046)**	-1.0022(0.3165)	0.0000(1.0000)	0.0000(1.0000)
312	**4.6032(0.0000)**	-0.7442(0.4570)	-0.0284(0.9774)	-0.0031(0.9976)	0.4027(0.6873)	0.2503(0.8025)	0.0319(0.9745)	0.0032(0.9975)
330	**-12.2383(0.0000)**	**9.3596(0.0000)**	-0.0218(0.9826)	0.0282(0.9775)	**5.7625(0.0000)**	**-4.4717(0.0000)**	0.1099(0.9125)	-0.0637(0.9492)
342	**-3.3867(0.0007)**	**2.1526(0.0317)**	0.0001(0.9999)	0.0120(0.9904)	1.4067(0.1598)	-1.4859(0.1376)	0.0297(0.9763)	-0.0198(0.9842)
364	-1.6626(0.0985)	-0.1433(0.8863)	0.0000(1.0000)	0.0000(1.0000)	0.1467(0.8836)	0.5395(0.5904)	0.0000(1.0000)	0.0000(1.0000)
370B	**-5.7705(0.0000)**	**3.4888(0.0006)**	-0.0161(0.9871)	0.0077(0.9938)	1.5150(0.1303)	-0.7974(0.4255)	-0.0112(0.9911)	0.0030(0.9976)
387	**-11.0832(0.0000)**	**9.0150(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	**11.2486(0.0000)**	**-9.4379(0.0000)**	0.0000(1.0000)	0.0000(1.0000)
412	**-6.7975(0.0000)**	**9.1065(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	1.0725(0.2837)	**-2.0415(0.0414)**	0.0000(1.0000)	0.0000(1.0000)
434	**-6.4145(0.0000)**	**4.2324(0.0000)**	0.0119(0.9905)	-0.0003(0.9998)	**2.4056(0.0163)**	-1.5959(0.1108)	0.0085(0.9932)	-0.0138(0.9890)
444	0.2784(0.7814)	1.3149(0.1922)	0.0000(1.0000)	0.0000(1.0000)	-0.2498(0.8032)	-1.1242(0.2635)	0.0000(1.0000)	0.0000(1.0000)
446	**-3.9405(0.0002)**	**2.5083(0.0142)**	0.0000(1.0000)	0.0000(1.0000)	0.4729(0.6371)	-0.6889(0.4922)	0.0000(1.0000)	0.0000(1.0000)
456	**-3.0049(0.0028)**	**2.2321(0.0260)**	0.0000(1.0000)	0.0000(1.0000)	1.1163(0.2646)	-1.4239(0.1548)	0.0000(1.0000)	0.0000(1.0000)
461	**-6.1325(0.0000)**	**2.6639(0.0078)**	0.0000(1.0000)	0.0000(1.0000)	**6.5424(0.0000)**	**-3.6579(0.0003)**	0.0000(1.0000)	0.0000(1.0000)
464	**-3.7809(0.0002)**	**6.3170(0.0000)**	0.0301(0.9760)	0.0136(0.9892)	1.1285(0.2595)	**-2.8537(0.0044)**	-0.0142(0.9887)	-0.0216(0.9828)
472	**-4.3592(0.0000)**	**5.2252(0.0000)**	-0.0116(0.9907)	-0.0130(0.9896)	-0.6934(0.4882)	**-3.6199(0.0003)**	0.0400(0.9681)	0.0138(0.9890)
490	**-4.6115(0.0000)**	**5.2286(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	1.3323(0.1831)	**-3.4744(0.0005)**	0.0000(1.0000)	0.0000(1.0000)
494	**-9.4181(0.0000)**	**6.1508(0.0000)**	0.0240(0.9809)	0.0095(0.9924)	**5.3143(0.0000)**	**-4.0853(0.0000)**	0.0039(0.9969)	-0.0215(0.9828)
503	0.1583(0.8743)	0.7654(0.4448)	0.0000(1.0000)	0.0000(1.0000)	-0.1162(0.9078)	-1.3217(0.1901)	0.0000(1.0000)	0.0000(1.0000)
505	**-6.2907(0.0000)**	**2.8215(0.0049)**	0.0000(1.0000)	0.0000(1.0000)	**3.5797(0.0004)**	-0.5974(0.5504)	0.0000(1.0000)	0.0000(1.0000)
533	**-8.6268(0.0000)**	**7.5132(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	**7.1249(0.0000)**	**-7.4010(0.0000)**	0.0000(1.0000)	0.0000(1.0000)
543	**-11.4055(0.0000)**	**11.9181(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	1.3584(0.1746)	**-5.2111(0.0000)**	0.0000(1.0000)	0.0000(1.0000)
548	-1.4671(0.1430)	**5.1451(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	**-3.5167(0.0005)**	-1.8622(0.0629)	0.0000(1.0000)	0.0000(1.0000)
556	**-6.1227(0.0000)**	**5.3186(0.0000)**	-0.0004(0.9997)	0.0132(0.9894)	1.5973(0.1105)	-1.8908(0.0589)	-0.0249(0.9801)	-0.0046(0.9963)
579	**2.7994(0.0052)**	1.0420(0.2977)	0.0018(0.9985)	0.0163(0.9870)	-0.4810(0.6306)	**-4.4804(0.0000)**	0.0327(0.9739)	-0.0263(0.9790)
595	**-3.2550(0.0013)**	**4.3907(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	1.4580(0.1457)	-1.6937(0.0911)	0.0000(1.0000)	0.0000(1.0000)
596	**-8.4528(0.0000)**	**5.0930(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	**4.5276(0.0000)**	**-2.5870(0.0099)**	0.0000(1.0000)	0.0000(1.0000)
604	**-2.5031(0.0128)**	**4.0659(0.0001)**	0.0000(1.0000)	0.0000(1.0000)	**2.3190(0.0208)**	**-3.2518(0.0012)**	0.0000(1.0000)	0.0000(1.0000)
625	-1.1267(0.2601)	**6.2904(0.0000)**	-0.0162(0.9870)	0.0212(0.9831)	**-4.6207(0.0000)**	0.3523(0.7247)	-0.0283(0.9775)	0.0045(0.9964)
647	-1.0040(0.3161)	1.9366(0.0537)	0.0000(1.0000)	0.0000(1.0000)	**-2.1428(0.0326)**	0.1873(0.8515)	0.0000(1.0000)	0.0000(1.0000)
657X	**-2.7878(0.0058)**	1.8180(0.0706)	0.0000(1.0000)	0.0000(1.0000)	**-4.6630(0.0000)**	**3.2341(0.0013)**	0.0000(1.0000)	0.0000(1.0000)
662	**-3.4592(0.0006)**	**3.4247(0.0007)**	0.0000(1.0000)	0.0000(1.0000)	1.1130(0.2663)	-0.6931(0.4886)	0.0000(1.0000)	0.0000(1.0000)
678	**-3.7498(0.0002)**	**4.7545(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	**3.6691(0.0003)**	**-4.9615(0.0000)**	0.0000(1.0000)	0.0000(1.0000)
703	-1.4841(0.1380)	1.9213(0.0549)	-0.0014(0.9989)	-0.0141(0.9887)	0.2717(0.7859)	-1.0219(0.3071)	0.0022(0.9982)	0.0144(0.9885)
741	**-2.5401(0.0114)**	**4.7584(0.0000)**	0.0000(1.0000)	0.0000(1.0000)	0.6921(0.4890)	**-3.9442(0.0001)**	0.0000(1.0000)	0.0000(1.0000)
748	**-8.8993(0.0000)**	1.8435(0.0654)	0.0498(0.9602)	-0.0149(0.9881)	**6.4175(0.0000)**	-0.5182(0.6044)	-0.0233(0.9814)	-0.0022(0.9982)
779	**-3.0884(0.0026)**	**3.4493(0.0008)**	-0.0125(0.9900)	0.0239(0.9809)	0.6400(0.5225)	**-2.0625(0.0397)**	0.0374(0.9701)	-0.0359(0.9714)
791	**-2.1069(0.0355)**	**3.4147(0.0007)**	0.0347(0.9723)	0.0030(0.9976)	1.8432(0.0659)	**-3.8796(0.0001)**	0.0033(0.9974)	-0.0219(0.9825)
794	**-4.6255(0.0000)**	1.2390(0.2156)	0.0000(1.0000)	0.0000(1.0000)	**3.3871(0.0007)**	-0.9399(0.3474)	0.0000(1.0000)	0.0000(1.0000)
811	**-6.1876(0.0000)**	**2.9515(0.0032)**	-0.0067(0.9946)	-0.0115(0.9909)	0.7026(0.4825)	0.8162(0.4146)	0.0094(0.9925)	0.0081(0.9935)
847	**-6.1494(0.0000)**	-0.5953(0.5518)	-0.0124(0.9901)	0.0152(0.9879)	**5.9144(0.0000)**	0.5409(0.5887)	0.0011(0.9991)	-0.0121(0.9904)
850	**-3.8875(0.0001)**	**-2.1499(0.0319)**	0.0158(0.9874)	0.0239(0.9809)	1.0377(0.2996)	**2.3230(0.0203)**	-0.0070(0.9944)	-0.0202(0.9839)
853	**-5.8508(0.0000)**	**3.8356(0.0001)**	-0.0134(0.9893)	0.0122(0.9903)	**2.0067(0.0450)**	**-2.2515(0.0246)**	-0.0149(0.9881)	-0.0009(0.9993)
863	**-4.6113(0.0000)**	**2.1579(0.0312)**	0.0000(1.0000)	0.0000(1.0000)	**4.1318(0.0000)**	**-3.0250(0.0025)**	0.0000(1.0000)	0.0000(1.0000)
888	**-2.1589(0.0316)**	**3.4320(0.0007)**	0.0857(0.9317)	-0.0792(0.9368)	1.3656(0.1730)	-1.9050(0.0576)	-0.0529(0.9579)	0.0378(0.9699)
893	-0.6347(0.5259)	0.9436(0.3458)	0.0000(1.0000)	0.0000(1.0000)	-0.6150(0.5387)	-0.2335(0.8154)	0.0000(1.0000)	0.0000(1.0000)
894	-1.4286(0.1538)	**2.0932(0.0369)**	0.0000(1.0000)	0.0000(1.0000)	0.6866(0.4928)	-1.5920(0.1123)	0.0000(1.0000)	0.0000(1.0000)
911	0.1362(0.8917)	**2.2036(0.0278)**	0.0188(0.9850)	-0.0085(0.9933)	**4.4813(0.0000)**	**-5.0808(0.0000)**	-0.0001(1.0000)	0.0029(0.9977)

Appendix C

Merger at the Word Level

Another advantage of using forced alignment to identify prenasal merger is to find out which PEN tokens that are the most and least frequently merged. Tables C.1 and C.2 show the results. We only consider words that have appeared more than 50 times in total and these results are based on all speakers. Identifying the most (or the least) frequently merged words at the individual level can also provide more characteristics for each speaker and potentially help with speaker identification.[1]

[1] However, one issue for us in using these results on speaker identification is that there are usually not enough tokens of the same words for each speaker for us to identify, except for some stopwords.

Table C.1: PEN Words with the Highest Proportions of Merged Instances.

Word	Merged Instances	Total Instances	Proportion of Merged Instances
pen	145	205	0.707
tennessee	141	213	0.662
ten	365	552	0.661
wednesday	35	54	0.648
hen	56	87	0.644
anyhow	37	58	0.638
anymore	113	183	0.617
anyway	193	326	0.592
men	141	239	0.590
many	563	956	0.589

Table C.2: PEN Words with the Lowest Proportions of Merged Instances.

Word	Merged Instances	Total Instances	Proportions of Merged Instances
september	33	140	0.236
suspenders	14	61	0.23
twenty	146	655	0.223
went	489	2201	0.222
whenever	51	251	0.203
cemetery	46	235	0.196
pencil	9	51	0.176
center	14	105	0.133
elementary	13	118	0.11
themselves	9	154	0.058

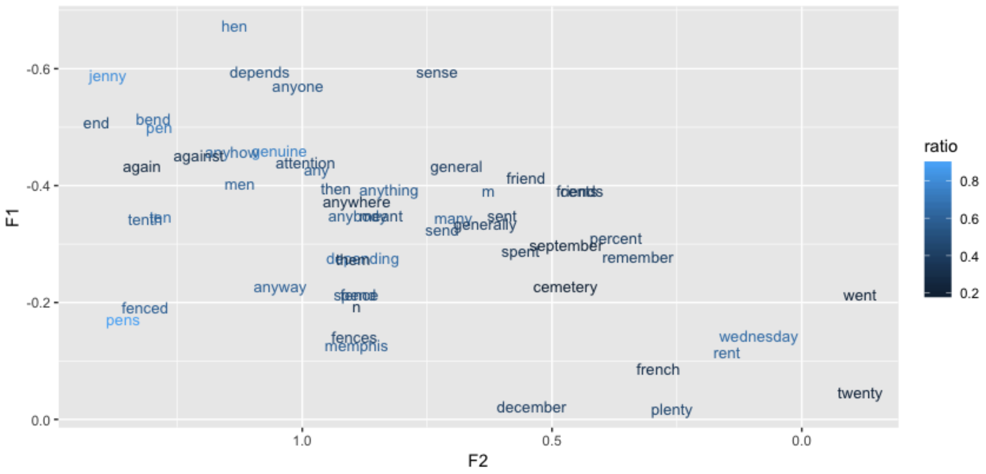


Figure C.1: Mean F1 and F2 Values of a Word in the Vowel Space.

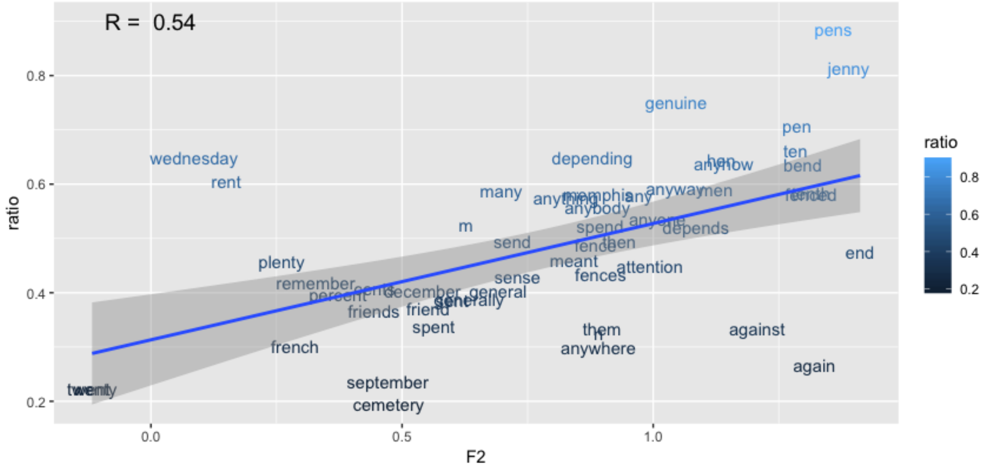
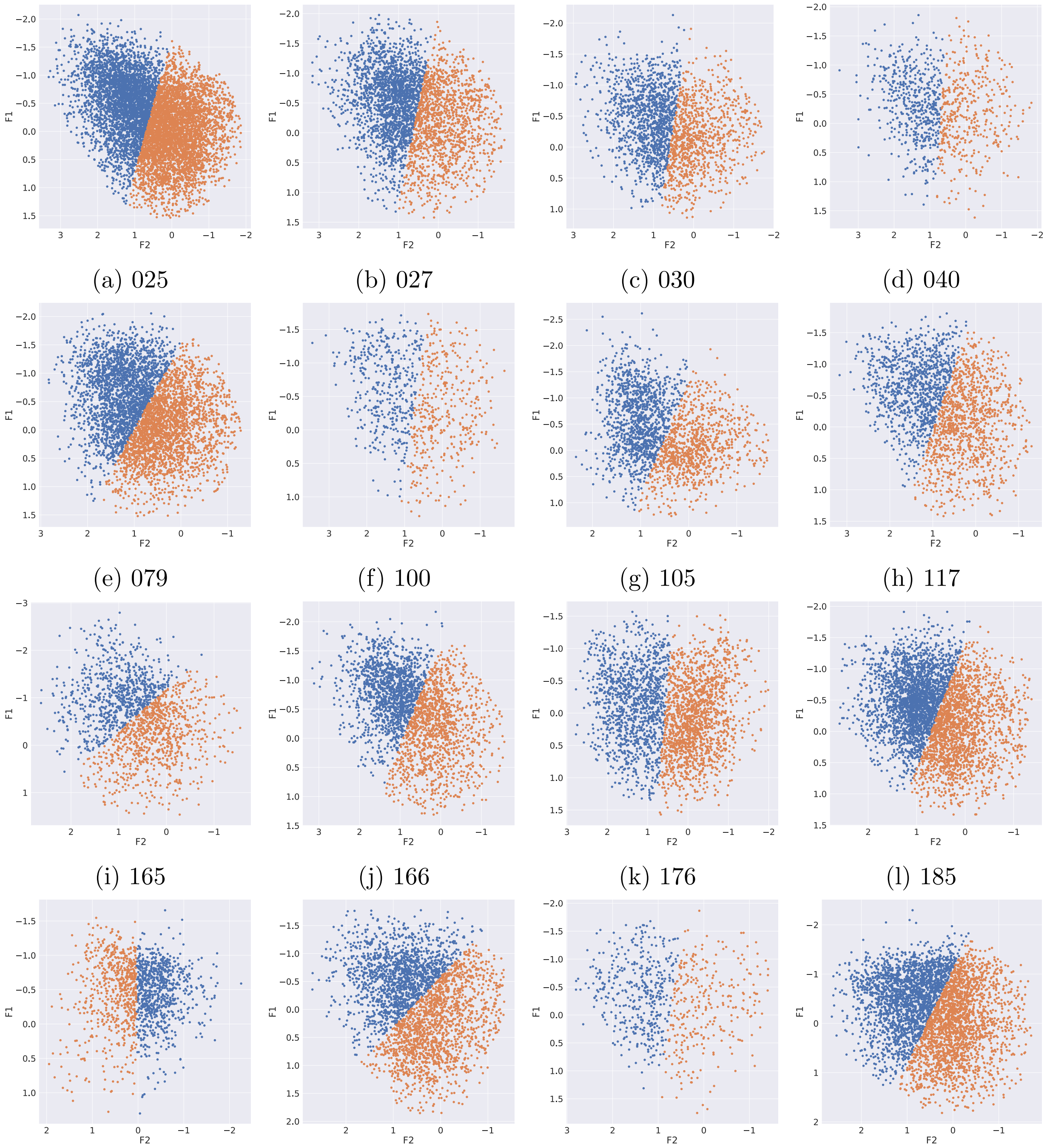


Figure C.2: Correlation between the Mean F2 Value of a Word and its Reclassification Rate.

Appendix D

Plot of K-means Clustering Results

In this chapter, we provide the plots of K-means Clustering (K = 2) on the vowel formant values for each DASS speaker. The caption of each following figure represents the speaker ID. The legends include two items: blue and orange represent labels 0 and 1 from K-means clustering. 0 should correspond to the vowel space of IH (low F1, high F2 values) and 1 should correspond to the vowel space of EH (high F1, low F2 values), but the plots of some speakers do not follow this pattern of data distributions. In total, we remove 8 out of the total 64 speakers whose data distributions do not follow the suggested pattern. And the IDs of these speakers are 252, 364, 444, 456, 472, 548, 779, 911.



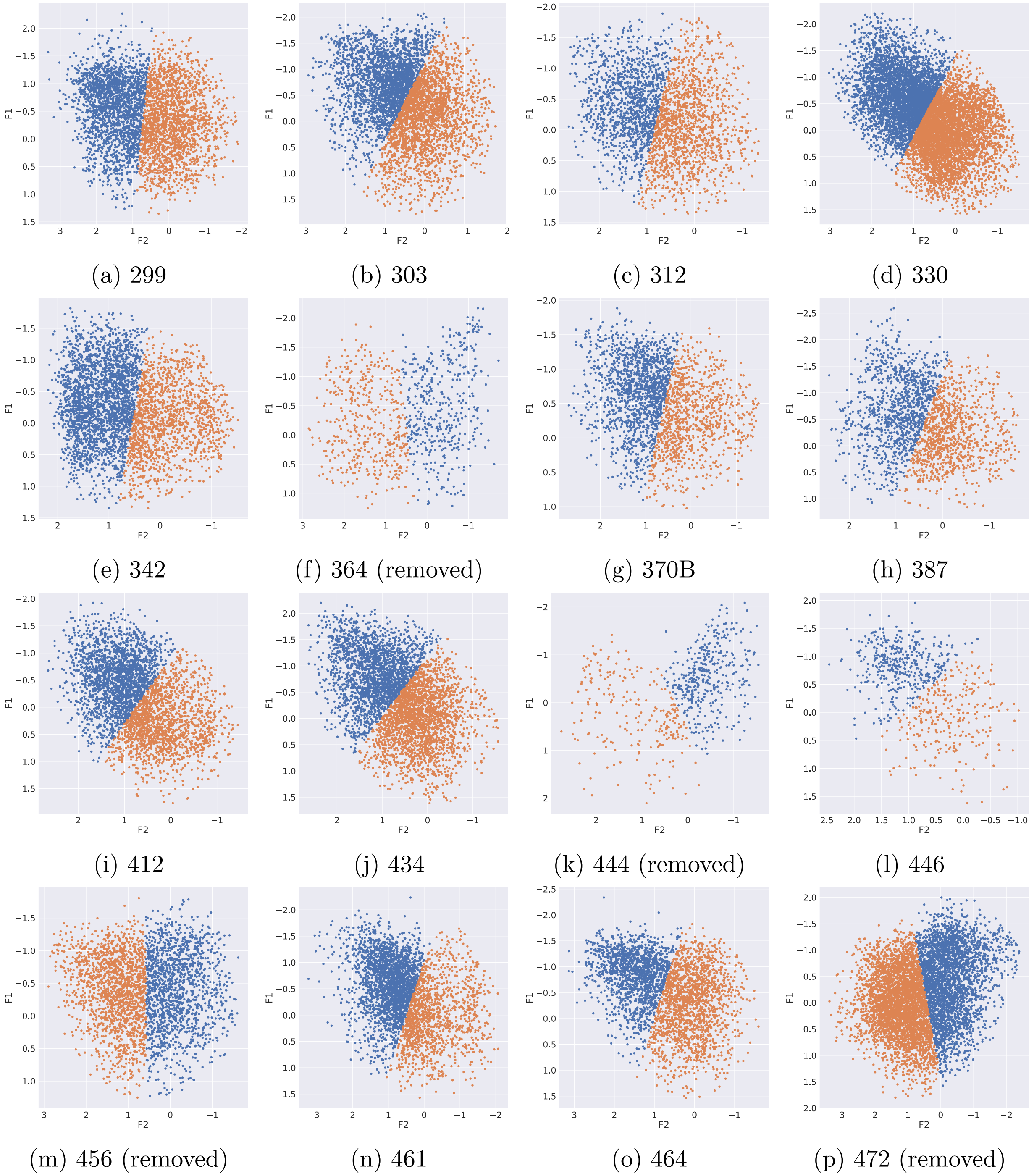
(a) 025 (b) 027 (c) 030 (d) 040

(e) 079 (f) 100 (g) 105 (h) 117

(i) 165 (j) 166 (k) 176 (l) 185

(m) 252 (removed) (n) 255 (o) 270 (p) 289

Figure D.1: 2-means Clustering Results for DASS Speakers 025-289. Blue and orange represent labels 0 and 1 from 2-means clustering.



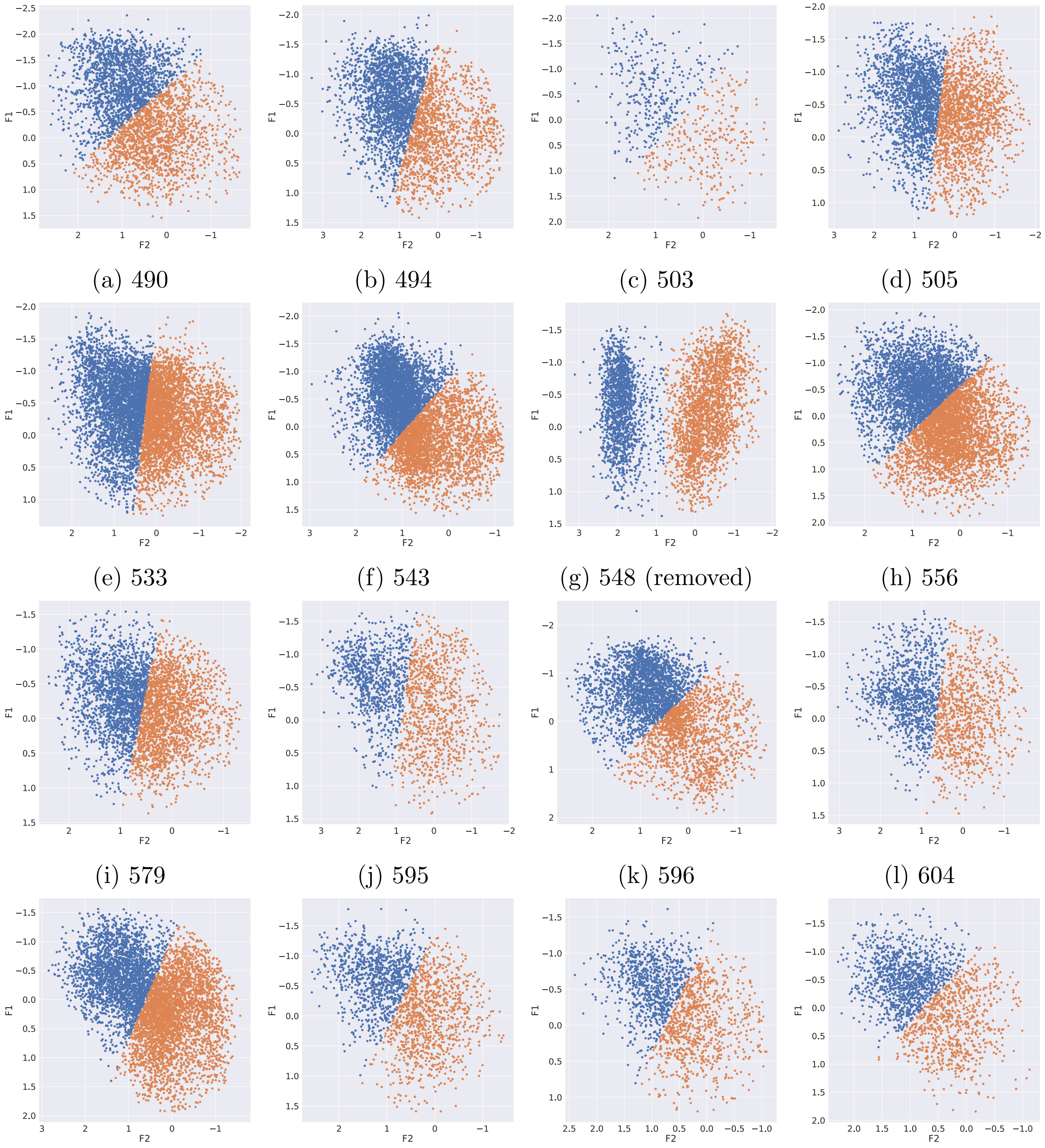
(a) 299 (b) 303 (c) 312 (d) 330

(e) 342 (f) 364 (removed) (g) 370B (h) 387

(i) 412 (j) 434 (k) 444 (removed) (l) 446

(m) 456 (removed) (n) 461 (o) 464 (p) 472 (removed)

Figure D.2: 2-means Clustering Results for DASS Speakers 299-472. Blue and orange represent labels 0 and 1 from 2-means clustering.



(a) 490

(b) 494

(c) 503

(d) 505

(e) 533

(f) 543

(g) 548 (removed)

(h) 556

(i) 579

(j) 595

(k) 596

(l) 604

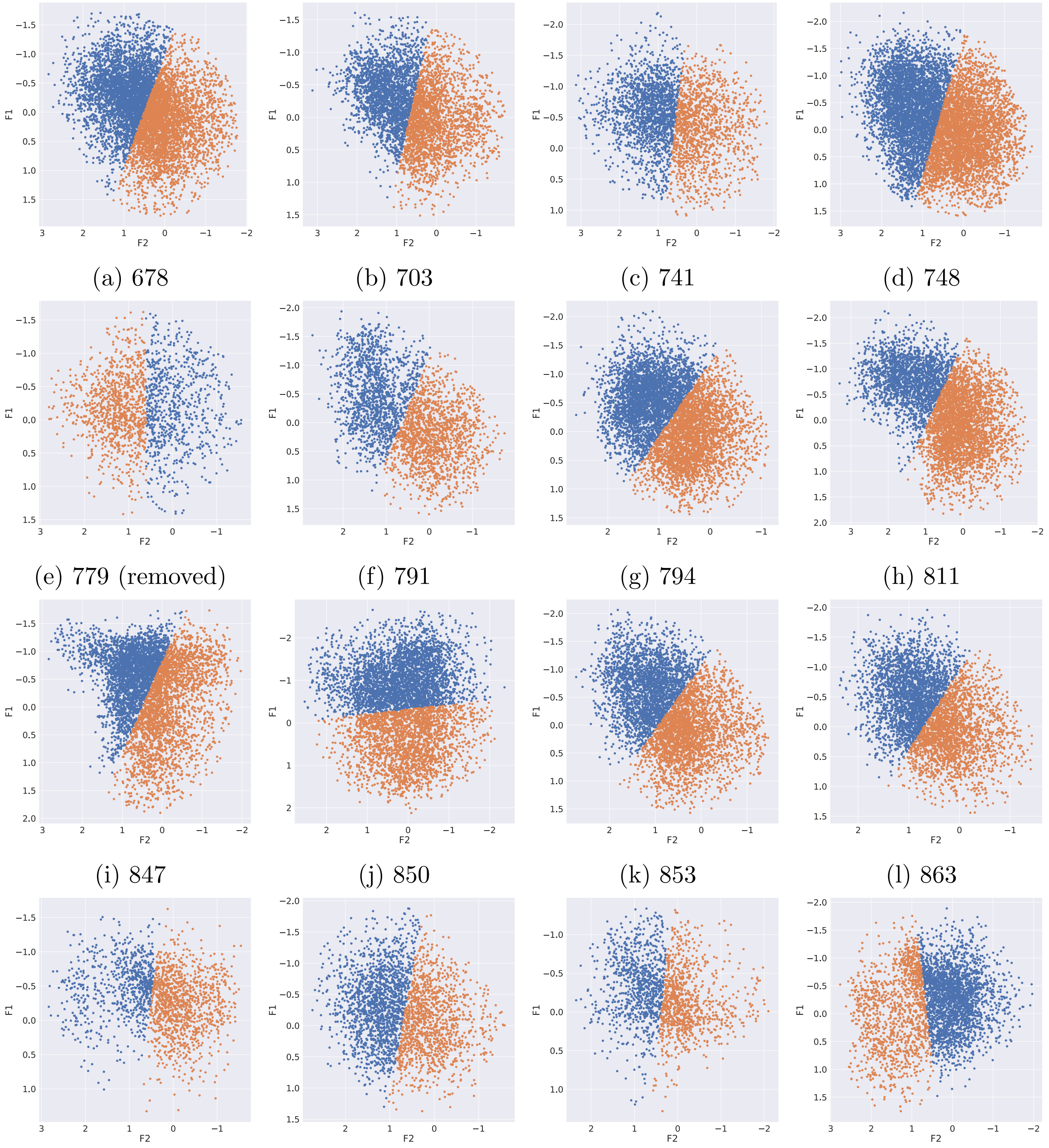
(m) 625

(n) 647

(o) 657X

(p) 662

Figure D.3: 2-means Clustering Results for DASS Speakers 490-662. Blue and orange represent labels 0 and 1 from 2-means clustering.



(a) 678

(b) 703

(c) 741

(d) 748

(e) 779 (removed)

(f) 791

(g) 794

(h) 811

(i) 847

(j) 850

(k) 853

(l) 863

(m) 888

(n) 893

(o) 894

(p) 911 (removed)

Figure D.4: 2-means Clustering Results for DASS Speakers 678-911. Blue and orange represent labels 0 and 1 from 2-means clustering.

Bibliography

[1] Vivian R Brown. "Evolution of the Merger of /ɪ/ and /ɛ/ before Nasals in Tennessee". In: *American Speech* (1991), pp. 303–315.

[2] Erik R Thomas. "Secrets revealed by Southern vowel shifting". In: *American Speech* 78.2 (2003), pp. 150–170.

[3] William Labov, Sharon Ash, and Charles Boberg. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter, 2008.

[4] Paul Boersma and David Weenink. *Praat: doing phonetics by computer*. [Computer program]. Version 6.0.56, retrieved 20 June 2019 from http://www.praat.org/.

[5] Jennifer Hay, Paul Warren, and Katie Drager. "Factors influencing speech perception in the context of a merger-in-progress". In: *Journal of phonetics* 34.4 (2006), pp. 458–484.

[6] Lauren Hall-Lew. "Improved representation of variance in measures of vowel merger". In: *Proceedings of Meetings on Acoustics 159ASA*. Vol. 9. ASA. 2010, p. 060002.

[7] Margaret E. L. Renwick and Rachel M Olsen. "Analyzing dialect variation in historical speech corpora". In: *The Journal of the Acoustical Society of America* 142 (2017), pp. 406–421.

[8] Shawn C. Foster, Joseph A. Stanley, and Margaret E.L. Renwick. “Vowel mergers in the American South”. In: *The Journal of the Acoustical Society of America* 142.4 (2017), pp. 2540–2540. DOI: 10.1121/1.5014282.

[9] Jennifer Nycz and Lauren Hall-Lew. “Best practices in measuring vowel merger”. In: *Proceedings of Meetings on Acoustics 166ASA*. Vol. 20. 1. ASA. 2013, p. 060008.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. “The Kaldi speech recognition toolkit”. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. 2011.

[11] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.” In: *Interspeech*. 2017, pp. 498–502.

[12] Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, and Jiahong Yuan. “FAVE (forced alignment and vowel extraction) program suite”. In: *URL http://fave. ling. upenn. edu* (2011).

[13] Keelan Evanini, Stephen Isard, and Mark Liberman. “Automatic formant extraction for sociolinguistic analysis of large corpora”. In: *Tenth Annual Conference of the International Speech Communication Association*. 2009.

[14] Jiahong Yuan and Mark Liberman. “Automatic detection of “g-dropping” in American English using forced alignment”. In: *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE. 2011, pp. 490–493.

[15] George Bailey. “Automatic detection of sociolinguistic variation using forced alignment”. In: *University of Pennsylvania Working Papers in Linguistics* 22.2 (2016), p. 3.

[16] Jiahong Yuan and Mark Liberman. "Investigating /l/ variation in English through forced alignment". In: *Tenth Annual Conference of the International Speech Communication Association*. 2009.

[17] William A. Kretzschmar Jr., Paulina Bounds, Jacqueline Hettel, Lee Pederson, Ilkka Juuso, Lisa Lena Opas-Hanninen, and Tapio Seppanen. "The Digital Archive of Southern Speech (DASS)." In: *Southern Journal of Linguistics* 37.2 (2013), pp. 17–38.

[18] Rachel M Olsen, Michael L Olsen, Joseph A Stanley, Margaret E. L. Renwick, and William Kretzschmar. "Methods for transcription and forced alignment of a legacy speech corpus". In: *Proceedings of Meetings on Acoustics 173EAA*. Vol. 30. 1. ASA. 2017, p. 060001.

[19] Cynthia G Clopper, David B Pisoni, and Kenneth De Jong. "Acoustic characteristics of the vowel systems of six regional varieties of American English". In: *The Journal of the Acoustical society of America* 118.3 (2005), pp. 1661–1676.

[20] Peter Ladefoged and Keith Johnson. *A course in phonetics*. Nelson Education, 2014.

[21] Joseph A. Stanley, William A. Kretzschmar, Margaret E.L. Renwick, Michael L. Olsen, and Rachel M. Olsen. *Gazetteer of Southern Vowels*. [Computer program]. Version 1.5. Retrieved July 2, 2019 from http://lap3.libs.uga.edu/u/jstanley/vowelcharts/. 2017.

[22] Erik R Thomas. *An acoustic analysis of vowel variation in New World English*. Vol. 85. Duke University Press Durham, NC, 2001.

[23] Adam Baker, Jeff Mielke, and Diana Archangeli. "More velar than/g: Consonant coarticulation as a cause of diphthongization". In: *Proceedings of the 26th West Coast conference on formal linguistics*. 2008, pp. 60–68.

[24] Jeff Mielke, Christopher Carignan, and Erik R Thomas. "The articulatory dynamics of pre-velar and pre-nasal/æ/-raising in English: An ultrasound study". In: *The Journal of the Acoustical Society of America* 142.1 (2017), pp. 332–349.

[25] Lee Pederson. “The Linguistic Atlas of the Gulf States: interim report four”. In: *American Speech* 56.4 (1981), pp. 243–259.

[26] Audacity Team. *Audacity: Free audio editor and recorder.* [Computer program]. 2014.

[27] Lisa Lipani, Yuanming Shi, Joshua McNeill, and Margaret E. Renwick. “Methods for noise reduction in a legacy speech corpus”. In: *The Journal of the Acoustical Society of America* 145.3 (2019), pp. 1932–1932.

[28] Daniel Jurafsky and James H Martin. *Speech and Language Processing, 2nd Edition.* Prentice Hall, 2008.

[29] Stanley S Stevens and John Volkmann. “The relation of pitch to frequency: A revised scale”. In: *The American Journal of Psychology* 53.3 (1940), pp. 329–353.

[30] Jiahong Yuan and Mark Liberman. “Speaker identification on the SCOTUS corpus”. In: *Acoustical Society of America Journal* 123 (2008), p. 3878.

[31] Mehryar Mohri, Fernando Pereira, and Michael Riley. “Speech recognition with weight -ed finite-state transducers”. In: *Springer Handbook of Speech Processing.* Springer, 2008, pp. 559–584.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. “Librispeech: an ASR corpus based on public domain audio books”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE. 2015, pp. 5206–5210.

[33] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. “The HTK book”. In: *Cambridge university engineering department* 3 (2002), p. 175.

[34] Peter Milne. “The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French”. PhD thesis. Université d’Ottawa/University of Ottawa, 2014.

[35] Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. “The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability”. In: *Speech Communication* 45.1 (2005), pp. 89–95.

[36] Josef R Novak, Nobuaki Minematsu, and Keikichi Hirose. “WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding”. In: *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing.* 2012, pp. 45–49.

[37] Prasanta Chandra Mahalanobis. “On the generalized distance in statistics”. In: vol. 2. National Institute of Science of India (Calcutta). 1936, pp. 49–66.

[38] Daniel Ezra Johnson. *Quantifying overlap with Bhattacharyya’s affinity.* 2015. URL: `https://danielezrajohnson.shinyapps.io/nwav_44`.

[39] Roy C. Snell and Fausto Milinazzo. “Formant location from LPC analysis data”. In: *IEEE Trans. Speech and Audio Processing* 1 (1993), pp. 129–134.

[40] Margaret E.L. Renwick and D. Robert Ladd. “Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts”. In: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7.1 (2016).

[41] Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, et al. “The subspace Gaussian mixture model A structured model for speech recognition”. In: *Computer Speech & Language* 25.2 (2011), pp. 404–439.

[42] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[43] William M Campbell, Douglas E Sturim, and Douglas A Reynolds. “Support vector machines using GMM supervectors for speaker verification”. In: *IEEE signal processing letters* 13.5 (2006), pp. 308–311.

[44] Jacob Goldberger and Hagai Aronowitz. “A distance measure between GMMs based on the unscented transform and its application to speaker recognition”. In: *Ninth European Conference on Speech Communication and Technology*. 2005.

[45] J-L Durrieu, J-Ph Thiran, and Finnian Kelly. “Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models”. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee. 2012, pp. 4833–4836.

[46] Shiyong Cui and Mihai Datcu. “Comparison of kullback-leibler divergence approximation methods between gaussian mixture models for satellite image retrieval”. In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2015, pp. 3719–3722.

[47] Tom Minka. *Divergence Measures and Message Passing*. Tech. rep. MSR-TR-2005-173. Jan. 2005, p. 17. URL: `https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/`.

[48] J. E. Shoup. “Phonological aspects of speech recognition”. In: *Trends in Speech Recognition*. Prentice-Hall, 1980, pp. 125–138.