A MULTIFACTORIAL, MULTITASK APPROACH TO AUTOMATED SPEAKER PROFILING

A dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Sean S. Simpson, M.S.

Washington, DC
September 5, 2019

Copyright © 2019 by Sean S. Simpson
All Rights Reserved

A MULTIFACTORIAL, MULTITASK APPROACH TO AUTOMATED SPEAKER PROFILING

Sean S. Simpson, M.S.

Advisors: Amir Zeldes, Ph.D. and Jennifer Nycz, Ph.D.

ABSTRACT

Automated Speaker Profiling (ASP) refers broadly to the computational prediction of speaker traits based on cues mined from the speech signal. Accurate prediction of such traits can have a wide variety of applications such as automating the collection of customer metadata, improving smart-speaker/voice-assistant interactions, narrowing down suspect pools in forensic situations, etc.

Approaches to ASP to date have primarily focused on single-task computational models– i.e. models which each predict one speaker trait in isolation. Recent work however has suggested that using a multi-task learning framework, in which a system learns to predict multiple related traits simultaneously, each trait-prediction task having access to the training signals of all other trait-prediction tasks, can increase classification accuracy along all trait axes considered.

Likewise, most work on ASP to date has focused primarily on acoustic cues as predictive features for speaker profiling. However, there is a wide range of evidence from the sociolinguistic literature that lexical and phonological cues may also be of use in predicting social characteristics of a given speaker. Recent work in the field

of author profiling has also demonstrated the utility of lexical features in predicting social information about authors of textual data, though few studies have investigated whether this carries over to spoken data.

In this dissertation I focus on prediction of five different social traits: sex, ethnicity, age, region, and education. Linguistic features from the acoustic, phonetic, and lexical realms are extracted from 60 second chunks of speech taken from the 2008 NIST SRE corpus and used to train several types of predictive models. Naive (majority class prediction) and informed (single-task neural network) models are trained to provide baseline predictions against which multi-task neural network models are evaluated. Feature importance experiments are performed in order to investigate which features and feature types are most useful for predicting which social traits.

Results presented in chapters 5-7 of this dissertation demonstrate that multi-task models consistently outperform single-task models, that models are most accurate when provided information from all three linguistic levels considered, and that lexical features as a group contribute substantially more predictive power than either phonetic or acoustic features.

INDEX WORDS: Automated Speaker Profiling, Computational linguistics, Sociolinguistics, Machine learning, Multi-task learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Automated Speaker Profiling (ASP) refers broadly to the computational extraction/prediction of speaker traits from the speech signal. The predicted traits may be physiological (e.g. height, weight, age, smoking habits), psychological (e.g. emotional state, stress level), or social (e.g. gender, ethnicity, education level, socio-economic status, dialect-region). Accurate prediction of such traits can have a wide variety of forensic, commercial, and medical applications such as narrowing down suspect pools, improving interactive voice-response systems, customizing service interactions, etc.

Nearly all ASP work performed on spoken data to date attempts to predict speaker characteristics via two basic steps. First, the speech signal is reduced to some type of vector-based representation (most often via the application of Gaussian Mixture Models [GMMs] to Mel-frequency cepstral coefficients [MFCCs] extracted from the signal). This vector representation is then used as a feature set for training a classificatory machine learning algorithm (often some type of Support Vector Machine [SVM]). In the majority of cases, the representation of the speech signal consists primarily of "low-level" acoustic information such as cepstral features, jitter, shimmer, sound pressure level, etc. and does not reflect much if any "high-level" information regarding specific phonetic, lexical, or discoursal phenomena. While such features may carry great explanatory weight for primarily biologically based speaker traits such as height and weight, they are less likely to be of use in predicting more socially-based traits such as dialect, education level, and to a lesser extent mixed biologically- and socially-based traits such as gender.

Despite the primary focus on low-level acoustic information in Automated Speaker Profiling using spoken data, there is a wealth of sociolinguistic literature demonstrating a link between usage patterns of phonetic, lexical, and discoursal variables and nearly every social trait one would desire an ASP system to predict. A number of authors for example have demonstrated links between realizations of certain phonetic variables and speaker gender (e.g. Eddington and Taylor, 2009; Stuart-Smith, 2007), age (e.g. Barbieri, 2008; Labov, 1966; Sankoff and Blondeau, 2007), ethnicity (e.g. Hoffman and Walker, 2010; Mendoza-Denton, 1997), and a host of other social categories. There is also a growing body of sociolinguistic work examining the social stratification of certain discourse and lexical phenomena (e.g. Cheshire, 2005; Johannsen et al., 2015; Tagliamonte and D'Arcy, 2007). In addition to the more traditional sociolinguistic work examining language features and social categories, there has also been a recent flurry of activity within computational sociolinguistics focused on the prediction of social categories from lexical and syntactic features present in text-based corpora– particularly corpora falling within the genre of Computer Mediated Communication (e.g. Ardehaly and Culotta, 2015; Bamman et al., 2014; Fink et al., 2012; Rao et al., 2010). Relatedly, work in the field of automatic speaker recognition (ASR) has also suggested that incorporation of such high-level features can improve accuracy and robustness in speaker discrimination tasks (e.g. Doddington, 2001; Reynolds et al., 2003; Kinnunen and Li, 2010). Such work has direct bearing on the present dissertation, though to my knowledge none of the resulting findings have been applied to spoken-language speaker profiling.

That the knowledge gleaned from sociolinguistic investigation of these socially-conditioned variables has not yet been applied to ASP efforts appears due more to limited inter-disciplinary communication rather than to any severe methodological or theoretical hurdles. This dissertation aims to bring sociolinguistic inquiry and Automated Speaker Profiling one step closer together by constructing a multi-factorial ASP system which takes into account current state-of-the-art approaches to classifying

speakers based on acoustic traits of the speech signal, as well as realization of certain phonetic, lexical, and discoursal variables known to pattern socially. It is expected that such a system will increase accuracy of speaker classification along speaker-trait axes which are largely or partially socially-based.

Finally, approaches to ASP to date have primarily focused on the prediction of one speaker-trait in isolation. Some recent work however has provided evidence that constructing a system which learns to predict multiple traits simultaneously, each trait-prediction task having access to the training signals of all other trait-prediction tasks, can increase the accuracy of speaker classification along all trait-axes considered (Poorjam et al., 2014; Weninger et al., 2012). This notion fits with the social theory of intersectionality– the idea that multiple overlapping aspects of social identity intersect to create a holistic individual social identity which is greater than the sum of its parts (Eckert, 1989; Levon, 2015). This concept has been borne out in variationist sociolinguistic studies as well– sociolinguistic variables have frequently been shown to be pushed one way by one aspect of identity yet pulled in a different direction by another. Labov's (1966) classic study of the social stratification of coda /ɹ/ in New York City for example demonstrated that realization of this variable was affected by a host of social factors, including gender, age, and socio-economic class. Because of the push and pull that various speaker traits may have on realizations of individual linguistic variables, this dissertation will take a 'Multi-Task Learning' approach (c.f. Caruna, 1997) to speaker profiling, allowing all trait-prediction tasks to be learned jointly, each essentially 'peeking' at what the others are doing.

1.1 Research Questions

Below I lay out the specific research questions I plan to address in this dissertation. These research questions are revisited in chapters 7 and 8 following the results and analysis presented in those chapters.

1. Does a multi-factorial system incorporating acoustic, phonological, syntactic, and lexical information result in significantly higher accuracy for speaker classification compared to systems which only take into account one type of linguistic information?

It's possible (though somewhat unlikely) that certain social traits may be wholly reliant on one or another type of linguistic feature. If this were the case, incorporation of features from other linguistic levels would not meaningfully improve classification accuracy, and a simple system including only that meaningful type would be preferred on the basis of computational and methodological efficiency.

2. If so, which type of cues hold the greatest explanatory weight for which social traits?

While there exist some assumptions floating throughout the sociolinguistic and speaker profiling literature that, for instance, phonological cues are likely more predictive of dialect region whereas lexical cues may be more predictive of education and social class (e.g. Jessen, 2007), there has not so far as I am aware been any kind high level, systematic investigation of just which sorts of cues are best suited to predict which sorts of social traits. Such an investigation would be useful for future work in speaker profiling.

3. Can a multi-task learning approach provide significant gains in accuracy over a system in which each speaker trait is predicted in isolation?

Recent approaches to ASP have demonstrated increased accuracy of speaker classification using multi-task learning approaches over systems which classify traits individually. This fits with the notion of intersectionality and the sociolinguistic understanding of how identity is performed linguistically, in that constellations of social traits interact and intersect to affect realizations of individual linguistic variables. This dissertation will follow recent work in attempting to predict each trait of focus first in isolation (single-task framework) and then in a multi-task framework to assess the relative accuracy improvements that a multi-task approach may provide.

1.2 Approach Overview

In this dissertation I focus on prediction of five different social traits: sex, ethnicity, age, regional origin, and level of education. Linguistic features from the acoustic, phonetic, and lexical realms are extracted from 60 second chunks of speech taken from the 2008 NIST Speaker Recognition Evaluation corpus and used to train several types of predictive models. Naive (majority class prediction) and informed (Single-Task Neural Network) models are trained to provide baseline predictions for each of the five social traits of focus against which the multi-task models are evaluated. Multi-Task models are then trained and compared to the naive and informed baseline models. Comparison of the multi-task models to the single task informed baseline models directly speaks to research question 3. Feature importance experiments are then performed on the best-performing multi-task models in order to address research questions 1 and 2.

1.3 Chapter Organization

The remaining chapters are organized as follows. Chapter 2 provides a literature review addressing the current state of Automatic Speaker Profiling, the sociolinguistic principles and theory on which this dissertation will rely, and the computational models and frameworks which will form the backbone of the speaker profiling systems I construct to address the above research questions. Chapter 3 provides an overview of the corpus that is used to train and test the speaker profiling models used in this dissertation, as well as a detailed explanation of how each feature was extracted and how the five social traits of focus are operationalized. Chapter 4 provides a high level exploration of the distribution of the various social traits examined throughout this dissertation with respect to the corpus, and an examination of how each extracted linguistic feature patterns with respect to these social traits within the corpus. Chapter 5 provides results of naive and informed single-task baseline experiments to predict the social traits examined in this dissertation, as well as all relevant methodological details regarding baseline model design, training, and evaluation. Chapter 6 provides the methodological details and results for the multi-task experiments predicting these social traits, and presents comparisons to the naive and informed baseline models. Chapter 7 examines the relative importance of various linguistic features and feature groups to the performance of each of the top-performing multi-task models for each social trait. Chapter 8 provides a detailed discussion of the differences between performance for single-task and multi-task models, and provides some actionable recommendations for deployment of these models in an automated speaker profiling context moving forward. Chapter 9 discusses the contributions made by this dissertation to the field of automated speaker profiling, as well as several limitations of the current work and potential improvements that could be made in the future.

Chapter 2: Background

This chapter presents a review of existing work and topics which have bearing on the present dissertation. I start by discussing previous work on automated speaker profiling, situating the computational approaches within the context of how human forensic profilers go about classifying unknown speakers demographically. I then turn to a discussion of the sociolinguistic principles that factor into this investigation, followed by a (somewhat) brief overview of the types of linguistic features known to be manipulated in the construction and expression of the aspects of social identity on which this dissertation will concentrate. Finally, this section concludes with a discussion of the machine learning tools and principles which form the basis of the speaker classification systems detailed in chapters 5 and 6.

2.1 Automated Speaker Profiling

Broadly put, the goal of Automated Speaker Profiling is to mine the speech signal in order to predict or extract certain social, physiological, or psychological/emotional characteristics of the speaker. As such, ASP systems are primarily employed in situations in which the identity of a speaker is unknown, but in which some type of demographic information about that speaker is useful to some end goal of the researcher.[1] Accu-

[1]There do exist applications of ASP in which speaker identity is known, such as the early detection of Parkinson's disease via analysis of acoustic cues (Bocklet et al., 2011), but I will not go into such applications in detail here.

rate prediction of such social/psychological/physiological traits can have a wide variety of forensic, commercial, and medical applications such as narrowing down suspect pools (Schilling and Marsters, 2015), automatically measuring customer satisfaction (Kamaruddin et al., 2016), remotely diagnosing health conditions such as obesity (Lee et al., 2013), etc.

Closely related to Automated Speaker Profiling is a sub-field sometimes termed Automated Author Profiling (AAP). Whereas speaker profiling attempts to predict characteristics of a speaker using spoken data, author profiling focuses on the prediction of characteristics of the authors of textual data. While the goals of ASP and AAP are the same, this difference in data-medium has led naturally to a difference in the features of focus for these two sister disciplines. Because of the textual nature of the data, researchers working in author profiling have tended to focus primarily on lexical, orthographic, and syntactic features in predicting author characteristics. Researchers working on speaker profiling on the other hand have focused almost exclusively on acoustic and phonetic features for use in predicting speaker characteristics. While great strides have been made examining the relative contribution of lexical and discoursal features to the accuracy of predicting social characteristics of authors from a wide range of textual genres, these findings have yet to be applied to profiling the 'authors' of spoken data. It is therefore unclear at this point to what extent features useful for profiling in textual genres are applicable to spoken genres, though the AAP literature represents a good repository from which to draw.

While computational acoustic-based approaches to speaker profiling have been on the rise since the early 2000's, the task of speaker classification has long been the province of human forensic experts using auditory and acoustic phonetic analysis (e.g. Jessen, 2007; Schilling and Marsters, 2015). To better understand the task computational systems are faced with in classifying speakers along trait-axes, it is useful to

examine the methods used by human profiling experts in performing such tasks.

2.1.1 Human forensic approaches to speaker profiling

Human speaker profiling takes place almost exclusively in the realm of forensic linguistics, a branch of applied linguistics particularly concerned with the application of linguistic knowledge to legal contexts (Coulthard et al., 2016).[2] As such the focus is typically on language as it relates to criminal investigations– narrowing down suspect pools, verifying the claimed identity of a speaker, etc. (Schilling and Marsters, 2015). Despite this particular legal focus, the same goals apply in human speaker profiling as in automated speaker profiling, namely the accurate prediction of certain characteristics of an individual based solely on his or her speech.

In contrast to most automated methods today, human expert forensic profilers typically attempt to use cues at all linguistic levels in speaker categorization (Schilling and Marsters, 2015). Furthermore, profiling experts also recognize that cues at different linguistic levels may be more informative than others for predicting certain traits. Phonetic and phonological cues may be particularly critical in profiling region of speaker origin for instance, whereas lexical and morphosyntactic cues may be more useful in approximating education level or occupation type (Jessen, 2007). Human profilers often start by first identifying the speaker characteristics that are important for the particular investigation at hand, and then consulting the relevant (socio)linguistic literature to determine which cues at the syntactic, phonetic, morphological, and lexical levels

[2]Though some may also consider (first-wave) sociolinguistics as a discipline engaged in human-driven speaker profiling, the sociolinguistic motivation and direction of inquiry is the direct inverse of that of speaker profiling. Sociolinguists investigate how and why aspects of social identity may influence speech, whereas speaker profiling investigates how speech may indicate aspects of social identity. Though findings from these two realms are undoubtedly related, sociolinguists are rarely if ever in the business of classifying unknown speakers along social trait axes, whereas this is the prime directive of forensic and automated speaker profiling research. Simply put, sociolinguistics and speaker profiling are two trains traveling opposite directions along parallel tracks.

have been shown to have bearing on said speaker characteristics. Once these features are established, the profiler will begin an auditory analysis paying specific attention to those linguistic features which have bearing on the investigation.[3] Depending on the objective of the investigation at hand, auditory analysis may be further augmented by acoustic phonetic analysis to examine the speech signal in more fine grained detail– particularly in cases in which fine-grained dialectal divisions are crucial.

This highlights a further distinction between human and automated methods of speaker profiling, namely the motivation behind what sorts of specific speech features are examined in making speaker classification determinations. Whereas much of the work on automated speaker profiling over the last decade has taken a buckshot approach to extracting a wide range of acoustic phonetic information from the speech signal, human profilers pay specific attention to those linguistic features known to vary (and the ways in which they vary) according to the speaker traits they are trying to uncover. Hill in his review of speaker classification concepts has this to say about focusing on features for which there is some empirical grounding:

> *"…if you wish to gather data relevant to classifying speakers, for whatever reason, you need to understand the attributes of speakers relevant to your required classification, rather than simply hoping that a genetic algorithm, neural net, Gaussian Mixture Model, or whatever will do the job for you. It might, but then again, it very well might not – at least, the classification will be nothing like as good as a properly informed discrimination that takes account of what you know about the populations of interest. … What we don't want to do is collect unstructured statistics in the hope that something will 'pop out' of the data."*

[3]It should be noted that while human profilers do pay specific attention to the cues relevant to the goals of the investigation at hand, such profilers are also usually expert linguists who are able to notice and take into account linguistic features which they may not have specifically been looking for, but which may nonetheless have bearing on their task.

— Hill (2007), pg.24-26

Focusing only on known features of course has the downside that such analyses will miss any cues important to speaker classification which have not been previously discussed in the literature, and it should be noted that current machine learning techniques if properly implemented can actually be quite proficient in determining which out of myriad features are important for classification and thus which to pay attention to and which to ignore. However, limiting the focus to an empirically grounded predictive feature set does increase computational efficiency and limit the degree of 'noise' feed to the system.

2.1.2 Standard approaches to automated speaker profiling

As mentioned in chapter 1, the vast majority of automated speaker profiling systems to date rely primarily if not entirely on acoustic features of the speech signal. Typically such systems rely on some combination of Mel Frequency Cepstral Coefficient (MFCC) statistics with other acoustic measures such as F0, jitter, voice quality, etc. extracted from the signal as predictive features for classification. MFCCs were initially introduced for use as features in Automatic Speech Recognition (ASR) in the 1980's, and remain heavily used in a wide variety of voice recognition/profiling applications today. MFCCs are the result of cosine transformation of the real logarithm of the short-term energy spectrum expressed on the Mel-frequency scale (Davis and Mermelstein, 1980). Roughly, MFCCs correlate with the shape of the vocal tract during speech production, thus making MFCC vectors excellent features for use in determining phone-identity over short time periods (hence their heavy usage in ASR). When averaged over longer segments of speech production, MFCCs can be of use in estimating the general physical attributes of a speaker's vocal tract, rather than determining which specific phone or

phone sequence is currently being uttered. However, as Hu et al. (2012) note, using MFCCs can have several drawbacks. First, MFCCs can drastically increase computational complexity depending on the length of the speech sample analyzed, since they capture linguistic information at very short time-scales (several ms). Second, MFCC measurements are greatly affected by recording environment, such that if training data is recorded with one microphone and testing data another, systems relying on MFCCs alone are unlikely to produce accurate results. This naturally limits the deployability of any system relying primarily on MFCC measurements for speaker classification.

Hybrid systems combining MFCC measurements with other acoustic features have performed quite well in predicting speaker traits which are wholly or partially biologically based. Profiling speaker sex operationalized as a binary classification task has been particularly amenable to such approaches. Levitan et al. (2016) for example using a logistic regression classifier trained on a combination of MFCC and F0 summary statistics report up to 95.2% accuracy in sex classification from short, 2 second snippets of recorded telephone speech. Similarly, Shafran et al. (2003) report 95.4% accuracy in sex detection using a HMM-based classifier trained on F0 and MFCC summary statistics over complete telephone-speech utterances. Hu et al. (2012) report some of the highest accuracy results in sex identification, achieving 98% accuracy using a two-stage classifier trained on several cleverly extracted F0 features and using MFCC features as a secondary gate, though it should be noted that the data-set used in this experiment consisted of extremely high quality laboratory recordings of speakers producing 77 digit sequences– not exactly realistic real-world data.

Efforts to estimate speaker age have had somewhat less success than efforts centered on sex/gender prediction, and, perhaps in part due to the increased difficulty of the task, there has been comparatively less work in this area of speaker profiling than on sex/gender prediction. One of the few attempts to computationally estimate linear

(i.e. numerical) speaker age is Poorjam et al. (2014). They report a Pearson correlation coefficient of 0.76 for males and 0.85 for females between actual and estimated linear age using Least Squares Support Vector Regression (LSSVR) based on i-vector transformation of MFCC features. More often researchers bin age into discrete categories, treating it as a categorical variable versus a linear one. Li et al. (2013) for example use a combination of prosodic and MFCC features to train a Support Vector Machine classifier in a four-way age classification task based on the Agender corpus (Burkhardt et al., 2010), achieving an unweighted accuracy of 45.8%. Similarly, Weninger et al. (2012) using a linear Support Vector Machine Classifier trained on the extended feature set (all acoustic features, but no MFCC features) of the 2012 INTERSPEECH speaker trait challenge (Schuller et al., 2012) report up to 61% unweighted average recall on the same four-way age classification task using the same data.

This subsection has been dedicated to discussing the basic standard approaches taken towards ASP to date. Below I turn to a discussion of some recent work that goes beyond this standard approach, taking into account features at levels beyond the acoustic.

2.1.3 Automated Speaker Profiling beyond the acoustic level

Compared to the plethora of work in automated speaker profiling which focuses on acoustic information as the basis for predictive features, there has been relatively little work examining the predictive power of features from other linguistic levels in the classification of speaker traits.

There have been a few recent attempts to incorporate phonetic information into ASP systems, but to the best of my knowledge such attempts have been entirely aimed at predicting speaker dialect. In these cases, separate GMMs are fit to MFCCs extracted

for each phone-type individually rather than for the speech signal as a whole. These phone-type vectors are then stacked into a super-vector prior to classification, effectively presenting the ML model with a summary-view of an individual's phonetic system (e.g. Biadsy, 2011; Biadsy et al., 2011). While effective for dialect classification, it should be noted that such an approach fails to take into account phonetic variables which may be contextually conditioned (e.g. fronting of /ŋ/ to /n/ in the progressive, pre-nasal raising of /æ/), multi-phonemic/reductive (e.g. consonant cluster reduction, word-final consonant deletion), or temporally dynamic (e.g. diphthong trajectories). It seems likely that, though perhaps computationally more expensive, the incorporation of such features may greatly improve predictive accuracy for dialect as well as any other social trait for which phonetic information may be useful. Though sociolinguistic evidence has been found for meaningful phonetic variation tied to gender (Stuart-Smith, 2007), age (Sankoff and Blondeau, 2007), ethnicity (Mendoza-Denton, 1997), and socioeconomic status (Trudgill, 1974), no ASP system which I'm aware of attempting to predict these categories has taken phonetic information into account.

Even less explored than phonetic information is the predictive power that discourse, (morpho)syntactic and lexical features might hold for spoken-language speaker profiling. While seldom used in speaker profiling, there is a wealth of evidence from author profiling studies that such features can be used to great effect in categorizing authors of textual data. Rao et al. (2010) for example used a number of lexical and orthographic features to classify Twitter users according to gender (male/female, 72% accuracy), age (above/below 30, 74% accuracy), regional origin (north/south India, 77% accuracy), and political orientation (Democrat/Republican, 83% accuracy). Nguyen et al. (2013) have even demonstrated that reasonable accuracy (micro F1 scores[4] between 0.85-0.87) can be achieved in classifying Twitter users by age-group and life-stage

[4]F1 score here refers to an evaluation metric ranging from 0 - 1 which is calculated based on the number of true positives, false positives, and false negatives observed. Micro F1 score is the weighted average F1 score over all classes.

based on unigram lexical information alone. Such work parallels sociolinguistic work demonstrating gender and age-based variation at the discourse, syntactic and lexical levels (e.g. Barbieri, 2008; Cheshire, 2005; Schleef, 2005), however the only work of which I'm aware that attempts to use such information in categorizing speakers of spoken data is (Sulayes, 2009; as reported in Schilling and Marsters, 2015), who worked with transcriptions of spoken data from the Switchboard corpus.

2.2 Sociolinguistic Foundations

In this section I go over a few sociolinguistic concepts key to this dissertation, and explain in more detail the motivation for taking a sociolinguistically grounded approach to automated speaker profiling.

2.2.1 The sociolinguistic variable

The atomic unit upon which variationist sociolinguistics is based is the "sociolinguistic variable." Before proceeding it is necessary to point out that I will be using this term in a somewhat expanded capacity compared to its traditional definition. The standard definition of a sociolinguistic variable is an underlying structure which has two or more identifiable surface variants which are referentially and semantically equivalent and not wholly dictated by surrounding linguistic structure but rather co-vary with social categories such as class, sex, and age. A canonical example of such a traditional sociolinguistic variable is the alternation between the alveolar (IN) and velar (ING) nasal in the progressive suffix "-ing" as in "running" (e.g. Trudgill, 1974). Because of the requirement for referential equivalence and identifiable surface variants, the application of the term "sociolinguistic variable" to morphosyntactic and lexical alternations which may not necessarily be referentially equivalent, or features which vary in terms

of presence vs. absence rather than variant vs. variant, is sometimes controversial.[5]

In this dissertation, I will be taking a broader view of the sociolinguistic variable, defining it for my purposes to mean any linguistic feature for which some social category X exhibits a demonstrably different usage pattern than some social category Y. Doing so allows the inclusion of features such as frequency of particular lexemes per thousand words under the umbrella of the sociolinguistic variable if such a feature indeed co-varies with some particular social category, whereas the traditional definition would exclude such a feature. It is my view that the difference between, say, two social groups evincing a difference in IN / (IN + ING) percentage and those same social groups evincing a difference in proportional usage of specific lexemes per thousand words lies solely in the operationalization of variant frequency, and that this is not a principled reason to distinguish them here.

2.2.2 Indexicality

Indexicality is the key principle through which sociolinguistic variables are theorized to be imbued with social meaning. In essence, indexicality refers to the ability for certain linguistic features to point to or "index" certain social categories and/or stances. These linguistic variables or features may take on different levels of indexical meaning depending on the social contexts in which they are used, or the level of conscious awareness speakers may have of them. Silverstein's (2003) "orders of indexicality" framework is typically the departure point from which indexical meaning is sociolinguistically treated at present. As operationalized by Johnstone and Kiesling (2008) and Eckert (2008a), 1st order indexical meaning is characterized by the correlation between a particular pattern of variant/feature usage and a particular socio-demographic group. This corre-

[5]see for instance Lavandera (1978) for an argument towards relaxing the requirement for referential/semantic equivalence, and Labov's (1978) subsequent rebuttal.

lation is one which an outsider could observe, but of which the users themselves are not necessarily conscious. First order indexicality roughly correlates with Labov's (1972b) concept of sociolinguistic "indicators"– features which may be associated with particular social groups but which do not exhibit patterns of stylistic variation and of which speakers are not consciously aware.[6] The transition from first order to second order indexical meaning comes to pass when differential usage patterns are in some sense recognized by speakers in the community, imbued with ideological meanings associated with and/or extracted from their underlying first order usage distributions, and so become available for social work (i.e. stylistic variation). Johnstone and Kiesling (2008) give the example of speakers in Texas drawing on a pre-existing division between urban and rural speakers in /ai/ monophthongization to imbue monophthongal /ai/ with associated meanings of ruralness, thus making /ai/ monophthongization available for social work as a linguistic resource for claiming a rural identity and "authentic Texanness." Second order indexicality roughly correlates with the Labovian concept of linguistic "markers"– those features which due to pre-existing usage differences within the community take on social significance and exhibit stylistic variation, but which are not typically overtly commented on (and not necessarily consciously recognized by their users). Second order indexicals may take on third order indexical meanings when they enter the conscious awareness of community members and thus become available for explicit meta-discursive comment, becoming manipulable at will for stylized performance of the identities or social characteristics associated with them. Third order indexicals correspond with what Labov termed the 'stereotype'– features which are overtly socially commented on and may become magnified in stylized performance such that they no longer reflect the form actually used by the community at large. Though in Silverstein's original framework there is theoretically no end to possible orders of indexicality,[7] so-

[6]See Johnstone and Kiesling (2008), pg.8-9 for a detailed comparison between the Labovian taxonomy and Silverstein's orders of indexicality.

[7]Silverstein discussed not first, second and third order indexicality, but rather nth and n+1st order indexicality, wherein the n+1st order indexicality comes about by assigning to the nth order some

ciolinguists operationalizing his framework typically do not attempt to define indexical orders above the third.

Indexicality is at the heart of variationist sociolinguistic research,[8] and is no less at issue in systems of automatic speaker profiling. To the extent that such systems are designed to recognize social divisions from speech, they are (or should be) in fact exclusively concerned with those speech features which hold some level of indexicality. However, as Eckert (2008a) points out, higher order indexical meanings are not as straightforward as they are sometimes considered to be. Features instead may carry a host of second and third order indexical meanings associated with them, the particular meaning drawn upon at a particular point in time only becoming clear within the specific context of the speech event. To deal with this reality, she introduces the concept of the indexical field: *"[a] constellation of ideologically related meanings, any one of which can be activated in the situated use of the variable"* (Eckert, 2008a, pg.454). In other words, one can't interpret the second order indexical meanings of a particular variant without reference to who said it, who they said it to, and the context in which it was said.

The issue of how to deal with and disambiguate higher orders of indexicality is potentially problematic for automated profiling systems seeking to take such linguistic features into account. In some ways however, the driving purpose of automated profiling systems may allow us to abstract away from some of the problems inherent in interpreting higher order indexical meanings. While sociolinguists are typically concerned with the question "why this now," for ASP systems it is largely irrelevant why a speaker uses a particular variant, so long as their usage pattern of that particular variant is correlated with some socio-demographic category. For instance, whether a

meta-pragmatic meaning according to a nativized ideological schema. In this framework it is possible to create n + 1 + 1 … + 1 orders of indexicality ad infinitum via successive meta-pragmatic shellacking.

[8]Though as Eckert (2008a) points out the distinction between 1st and 2nd order indexicality is not always respected by variationists.

speaker uses high proportions of the velar ING variant to index formality, articulateness or intelligence within a given speech sample is irrelevant if the speaker is an educated female and the system has been trained to recognize that high ING/IN proportions in concert with other observed features are indicative of educated females. That is to say, the exact stance indicated by a variant at any given time is irrelevant so long as the constellation of stances typically indicated by that particular variant usage pattern are statistically correlated with sociodemographic categories. In other words, so long as the variant maintains some socio-demographic distinction in baseline usage patterns, the higher order indexical meanings inherent in its use are immaterial to the profiling system. This highlights the fact that, as ASP systems are primarily concerned with predicting macro-social demographic information,[9] they are more interested in first order indexicals than any specific accompanying higher order indexical meanings a particular variant may take on in any given situated usage.

This is not to say however that higher order indexical meanings are totally irrelevant to automated profiling systems. Insofar as higher order indexical meanings associated with a variable vary group to group, and these meanings have bearing on demographic category prediction, it is important for such systems to take them into account. For instance, it may be desirable for a profiling system to have the ability to weight ING/IN ratio differently when attempting to predict the education level of a northerner vs. a southerner, or to weight the rate of expletives differently when gauging the age of a female vs. a male. In other words, while it is beyond the scope of such systems to take into account who the speaker was talking to and the pragmatic context of the utterance, it is possible to take into account the "context" of the speaker him/herself, and thus get at a more nuanced view of what higher order indexicals may point to on a case by case basis. This notion dovetails with the intersectional nature

[9]E.g. predicting speaker sex writ large vs. predicting performance of masculine/feminine gender roles to greater or lesser degrees within a particular conversation.

of identity, discussed in detail in section 2.2.3, and is one of the primary motivations for taking a Multi-Task Learning approach to speaker profiling in this dissertation, as detailed in section 2.4.2.

2.2.3 Intersectionality

Intersectionality as a social theory was first introduced in the 1980's and 1990's by black feminist scholars and others investigating gender and ethnic divisions primarily within the realm of sociology (e.g. Crenshaw, 1989; Hooks, 1981), and has since become an important part of the way in which social scientists across a host of related disciplines conceptualize and approach matters related to social identity. The basic tenet of intersectionality theory is that socially relevant categories (i.e. aspects of social identity) are mutually constitutive– in other words, gender, class, ethnicity, etc. all interact with and influence one another on a continual basis, and consequently no aspect of social identity is formed or expressed in isolation. Social identity writ large within an intersectional framework therefore is better considered as a synergistic, holistic web of aspects of identity which may smear into and modulate one another rather than simply an additive accretion of individual, isolated aspects (i.e. the whole is greater than the sum of its parts). For example, what it means to be a black female in America cannot simply be broken down into meanings associated with blackness, femaleness, and Americanness (Crenshaw, 1989). Likewise, what it means to be an American male in an urban community vs. a rural community is not simply the subtraction and addition of ruralness and urbanness– rather the concept and expression of masculinity itself along with its associated meanings may change subtly (or not so subtly) in combination with these and other aspects of identity (see e.g. Campbell et al., 2006).

One of the primary benefits of considering identity from an intersectional perspective within sociolinguistics is the recognition that language practices as they relate to social identity cannot be adequately considered with reference to a single aspect of identity alone. Sociolinguists working on sexuality and gender for example have been particularly vocal in calling for the treatment of gender to be contextualized against the broader backdrop of race, class, socio-economic status, and other social pressures at work within the community of study (Bucholtz, 1999; Eckert, 1989; Kirkham, 2015; Levon, 2015). The intersectional nature of identity has particular bearing on investigations of language variation, as it is rare that a linguistic feature is affected by a single aspect of identity alone, or that the effect of broad macro-social categories on a particular linguistic variable will be the same community to community or population to population. Linguists investigating variation have found over and over again that linguistic variables tend to be simultaneously affected by a host of social features (e.g. Eckert, 1989; Labov, 1966; Trudgill, 1974, among many others), and that markers of a particular aspect of social identity in one community may be not be used to mark that same aspect of identity in another community (e.g. Podesva and Van Hofwegen, 2014). In other words, as mentioned in section 2.2.2 above, the higher order indexical meanings associated with particular linguistic features may vary depending on the backdrop of the individual speaker's socio-demography and sociolinguistic experience. It was partially the recognition of the intersectional nature of identity and its effects on language variation that gave rise to the push for more ethnographically grounded, locally centered "second wave" and "third wave" (Eckert, 2012) studies within variationist sociolinguistics.

Insofar as automated speaker profiling systems attempt to predict or uncover the compositional social identity of a speaker, a responsibly constructed system ought to take insight from and be constructed in reference to prevalent theories as to how social identity is constructed/manifested in the first place. Such a grounding is not only

desirable from a theoretical perspective, but from a practical one as well. Taking into account the intersectional nature of identity and the multi-directional push/pull that various aspects of identity may have on individual linguistic variables, it becomes clear that a modular approach in which each trait is predicted independently is sub-optimal. Such a system has for example no way of tuning its expectations for how gender might be differentially expressed in the context of a southern speaker vs. a northern speaker if the gender and dialect prediction tasks have no communication. A system which allows each trait prediction task to "peek" at what other tasks are doing however, subsequently updating its own prediction in light of what it finds, would be able to perform such an operation– likely boosting accuracy in addition to bringing the system in line with current social theory regarding identity. The intersectional nature of identity is the primary motivator for taking a Multi-Task Learning approach to speaker classification in this dissertation– a framework which will be discussed further in section 2.4.

2.2.4 Disambiguating style and sociodemography

As this dissertation is focused on the differentiation of speakers along various social axes, it is primarily concerned with inter-speaker variation– i.e. linguistic cues and cue combinations which reliably distinguish coherent groups of speakers. A complication to this type of work however is that speakers themselves are not always internally consistent. They may vary in the realization of certain features from situation to situation, and indeed moment to moment. Enter the concept of sociolinguistic style and stylistic variation.

The traditional Labovian view of sociolinguistic style posits that intra-speaker variation is driven by the amount of attention one pays to one's speech as well as the perceived formality of a given speaking context (Labov, 1966, 1972c). Within such an

'attention to speech' model, higher rates of self-monitoring and greater attention paid to speech are typically associated with an increased likelihood of using standard, overtly prestigious variants over non-standard variants, whereas lower rates of self-monitoring and less attention paid to speech are associated with a decreased likelihood of using standard over non-standard variants. Further work has demonstrated that intra-speaker stylistic variation may, in addition to attention paid to speech, be conditioned by such factors as topic of conversation (Rickford and McNair-Knox, 1994) and stance-taking (Bucholtz, 2009; Kiesling, 2004), among others, and that it may be quite rapid and fine grained (Coupland, 1980).

The phenomenon of style shifting presents a problem to those systems attempting to predict socio-demographic categories based on short speech samples, in that many of the linguistic features which vary during style-shifting are co-linear with features which may be useful in distinguishing demographic categories from one another.[10] That is to say, if style is not properly controlled for, how could one tell whether, for instance, a speaker's high incidence of non-standard variant usage was due to using a particularly relaxed and informal style, rather than some aspect of socio-demography which we might also expect to be associated with high proportions of non-standard variants. in short, if speakers within the training and/or testing data are using different stylistic registers, style becomes a confounding variable in any analysis seeking to distinguish sociodemographic categories from one another.[11] In order to properly train a system to detect feature differences and distinguish coherent sociodemographic groups on the basis of these linguistic differences, clearly it is necessary to control for style as much as possible.

[10]This phenomenon is in fact so pervasive and was so problematic for first-wave variationist research that it served as the primary motivation for the development of Bell's (1984) influential framework of Audience Design.

[11]Labov (1972b) concisely characterizes this problem as the difficulty in distinguishing between the speech of a casual salesman and a careful pipe-fitter.

Perhaps the most obvious way to control for style is to control the speech event from which data in the test and training corpora are drawn. Speech samples within the corpora should, as much as possible, be drawn from the same genre of speech event, collected in similar settings, and come from speakers addressing similar interlocutors. The NIST Speaker Recognition Evaluation test/training corpora which is used for training and testing the ASP systems constructed during the course of this dissertation (discussed in detail in chapter 3) fit these criteria quite well in most accounts. First, all speech samples included in the NIST SRE sets come from either conversational telephone speech or conversational interview speech, allowing for control of speech event genre. In addition, all interlocutors within the NIST SRE sets are strangers to one another, eliminating any confounding effect of variable interlocutor familiarity. The NIST SRE sets do leave something to be desired however, as the interlocutors with whom the speaker of a speech sample was talking were not kept constant (although they were consistently strangers to one another). This introduces a confound of speech accommodation, in that speech samples from two speakers who are demographically identical may have been drawn from conversations with interlocutors who differ greatly in their variant usage, prompting different levels of accommodation and therefore different linguistic feature configurations from two speakers who might otherwise be expected to behave quite similarly in terms of linguistic feature configuration. Furthermore, the speech events from which the NIST SRE data were drawn were not controlled for topic, introducing another potential confounding variable.

Aside from controlling the speech event as far as possible (though it must be recognized that it is relatively impossible to control for all potential confounding factors between two speech events), a further strategy for mitigating the confounding influence of style shifting is to train the system on a large volume of data. The more data the system is trained on, the lower the effect of any random noise introduced via style-shifting in any one particular speech sample. Again the NIST SRE data sets used

for training and testing the ASP models developed during this dissertation satisfy this criteria, comprising 1,169 five minute telephone conversation recordings (97.4 hours of speech in total) from 454 unique speakers.

2.3 Sociodemographic Categories of Focus

This dissertation attempts to classify speakers along the following sociodemographic axes:

1. Sex
2. Age
3. Region
4. Ethnicity
5. Education

While traits 1 and 2 have both a social and a biological component, traits 3-5 are wholly social. The first three traits have been extensively examined in previous ASP studies, enabling comparison of the current methodology to previous work. Traits 4 & 5 however have, to the best of my knowledge, only been incorporated into work on ASP by Gillick (2010), who used exclusively lexical features as predictors. Investigation into which types of features contribute most to accuracy of speaker classification along traits 4 & 5 thus represent a novel contribution of this dissertation to the field of automated speaker profiling (though work has been done on classifying authors according to these traits using text-based corpora).

What follows is a short overview of each of the sociodemographic traits of focus identified above, addressing some of the issues inherent in their operationalization in an ASP system, and briefly touching on the relevant sociolinguistic literature. For an exhaustive list of the predictive feature set used to classify speakers in this dissertation, see chapter 3.

2.3.1 Sex

Gender is perhaps the most widely studied of all social traits within sociolinguistics other than region. While in most "first wave" early variationist work gender was treated as a binary variable coinciding with biological sex, within the last few decades researchers working on gender and language have made a push for a contextualization of this variable, recognizing that gender is not a property one "has," but rather a thing one "does," and thus not entirely coincidental with biological sex (e.g. Eckert and McConnell-Ginet, 1999, 2003). Likewise there has been a flurry of recent work investigating language in relation to transgender individuals in a non-binary context (e.g. Brown, 2011, 2015; Hazenberg, 2012). While recognizing the complex nature of gender, ASP systems in practice are typically aimed at the prediction of binary gender.[12] For this reason, this subsection focuses specifically on sociolinguistic variables as they relate to the binary expression of speaker sex.[13]

It is important to note here that while many of the (particularly acoustic) linguistic variables studied in relation to gender have a biological basis for differentiation, sociolinguistic investigation into the usage of these variables has often shown that their realization is manipulable as a resource in performing gender identity. Thus, the linguistic expression of sex is at once both biologically and socially based. Fundamental Frequency (F0) is a prime example of this. Because males typically have thicker and longer vocal folds than do females (Simpson, 2009, pg. 622), and the male larynx is approximately 50% larger on average than the female larynx (Podesva and Kajino, 2014, pg. 104), males' vocal folds tend to vibrate more slowly during speech than do females',

[12]Though the prediction of non-binary gender and sexual orientation is an interesting area for future research in this field

[13]I use the term "sex" rather than "gender" throughout this dissertation when referring to participants of the NIST SRE 2008 corpus as this is the term present in the corpus metadata and presumably therefore the term used by corpus collectors in asking speakers to identify themselves with respect to this category.

resulting in a lower average fundamental frequency. While there is a biological basis for sex differentiation in average F0, many studies have provided evidence that these biological differences have been mapped on to stereotypical conceptions of masculinity and femininity (i.e. have taken on higher order indexical meanings), above and beyond biology, and that consequently F0 may be drawn upon as a linguistic resource by speakers in constructing and performing (gender) identity. For instance, the magnitude of difference in fundamental frequency between men and women has been shown to vary cross-culturally, despite a lack of corresponding cross-cultural variation in the magnitude of sexual dimorphism between males and females. This indicates that speakers in different cultures may enhance or minimize the natural F0 difference between men and women to varying degrees. Yuasa (2008) for example found that the difference in average F0 for male and female Japanese speakers is significantly wider than the range for male and female American English speakers (mostly due to rather high average F0 values for Japanese females— which Yuasa hypothesizes may result from the inordinately high esteem which Japanese women place on femininity). Differences in F0 between sexes have also been shown to exist in prepubescent children, despite the fact that the sex-based physiological differences typically linked to difference in F0 don't emerge until after puberty (e.g. Graddol and Swann, 1983). Findings like these represent solid evidence that, even in the absence of any meaningful physical difference, F0 can be manipulated to invoke and reproduce the dominant gender stereotypes of the community.

A number of studies have investigated phonetic correlates to speaker gender. Often for binary phonetic variables, females will tend to exhibit higher ratios of the 'standard' or 'overtly prestigious' variant. Labov's (1966) classic study of New York City English for example found that females tended to realize the standard rhotic variant of coda-/ɹ/ and the standard non-stopped variants of th/dh more frequently than males in many social classes. Likewise Trudgill's (1974) study of Norwich English showed

females with consistently higher usage of the standard velar variant of the ING variable than men across all social classes. Not only do females tend to use higher ratios of standard variants, they also have been shown quite consistently to lead incipient sound changes taking place within communities. For instance, females tend to have more progressive vowel realizations in most of the major vowel shifts in the United States described in the literature (e.g. Baranowski, 2008; Hall-Lew, 2005; Kennedy and Grama, 2012; Labov, 2001; Ward, 2003). Given this, some metric of standard/non-standard phonetic variant ratio and the degree of advancement of known sound changes may be useful in an automatic system attempting to predict speaker gender. There is also evidence that, in addition to leading vowel shifts, females tend on average to produce longer vowels and have larger vowel spaces than do males (Munson, 2007; Neel, 2008; Simpson and Ericsdotter, 2007), suggesting that some measurement of vowel dispersion and average duration may also be useful in automated speaker profiling.

In addition to differences in the phonetic realm, there is a growing body of research investigating gender-based differences at the lexico-syntactic level. Cheshire (2005) for example describes robust gender differences in the way that discourse-new entities are syntactically flagged for adolescents, and a number of studies have pointed to gender-based differences within the quotative system (e.g. Barbieri, 2007; Blyth et al., 1990; Tagliamonte and D'Arcy, 2004). Likewise Mondorf (2002) describes significant gender-based variation in the use of finite adverbial clauses, causal clauses, postposed conditional clauses, purpose clauses, and concessive clauses in addition to an overall female preference for postposed clauses and a male preference for preposed clauses. Recent computational and corpus work has also revealed substantial differences in the frequency with which males and females use different parts of speech. Johannsen et al. (2015) for example demonstrates a correlation between males and the use of nouns, impersonal pronouns, and numerals, and a similar correlation between females and higher rates of personal pronoun usage. They also found correlations between gender and

preference for certain universal dependency relations.

Evidence for gender-based variation at the discourse level seems somewhat mixed. Though early researchers postulated gender differences for things like degree of hedging (e.g. Lakoff, 1975) and positive/negative affective stance (e.g. Anderson and Leaper, 1998), more recent corpus work suggests that these earlier findings largely do not hold or were likely the product of topic variation which was not properly controlled for (Bayard and Krishnayya, 2001; Precht, 2008; Thomson, 2006). One of the stereotypes about men's and women's language at the discourse/lexical level that does appear to hold however is the higher use of expletives by men (Precht, 2008).

2.3.2 Age

Throughout the sociolinguistic literature, age is largely treated as a mechanism for detecting language change over apparent time. Relatively few studies have examined age from the point of view of a sociolinguistic variable in its own right (Barbieri, 2008; Eckert, 1997), though researchers have long noted that diachronically stable sociolinguistic variables may exhibit age-grading across the lifespan (Labov, 1994). From the point of view of an ASP system, it's largely irrelevant whether differences seen across age groups are the result of diachronic language change or stable age-grading, so long as the differences reliably separate age groups (and contemporaneous training data is available).

More pressing is the question of how to operationalize age. In sociolinguistic studies of age, researchers typically bin speakers into discrete age-groups (e.g. 15-30, 30-45, etc.). One problem with this binning technique is that the number of bins varies widely from study to study, making cross-study comparisons difficult. Furthermore, the motivation behind the construction of these bins is not often clear. Bins may

attempt to roughly delineate generations, life-stages, or may simply be dictated by the idiosyncrasies of the data at hand. Treating age as a linear variable also has its challenges, as often a researcher will either have not enough data or not enough dispersion within the data to merit such a treatment. Further complicating the matter is that, as Eckert points out, *"…chronological age can only provide an approximate measure of the speaker's age-related place in society"* (1997, pg. 155). For this reason Eckert and others have suggested that life-stage rather than biological age should be taken as the locus of age-based variation (though the operationalization of 'life-stage' is no less fraught). Because measures of life-stage are not typically available in large, spoken-data corpora, my treatment of age in this dissertation will focus on chronological rather than social age. Details on how age is operationalized within this dissertation are available in chapter 3.

Much of the evidence for phonetic and phonological age-based variation in American English comes from studies investigating changes in progress. Most of this work centers around shifts within the vowel system. For instance, younger speakers have been shown to exhibit more fronted back-vowels (Hall-Lew, 2011) and more retracted and lowered short front vowels (Kennedy and Grama, 2012) in the California Vowel Shift. Likewise younger speakers lead the fronting of back vowels in the Southern Vowel shift, along with the near reversal of the tense and lax front vowels (Fridland, 2001). Younger speakers also lead the raising of /æ/ and all accompanying vowel re-arrangements described by Labov et al. (2006) as the Northern Cities Shift.[14] Jacewicz et al. (2011) describe a more general sound change across the United States wherein younger speakers lead the lowering and retracting of the short front vowels /ɪ, ɛ, æ/, which they dub the North American Shift. It appears that in concert with dialect-region identification, the relative position of back and short front vowels may be an important indicator of speaker age in the profiling systems constructed in this dissertation.

[14]Though some recent work has suggested this trend may be reversing. See e.g. Wagner et al. (2016)

Investigation into stable age-grading also provides insight into what phonetic variables to pay attention to in predicting speaker age. Labov (2001) for example has demonstrated that the stable ING variable (discussed above) tends to show a curvilinear pattern across life-stages, with usage of the non-standard IN variant peaking in late adolescence and then gradually retreating as speakers enter the workforce. Chambers' (2003) discussion of 'retrenchment' suggests that this pattern of post-adolescent back-off (potentially followed by post-retirement resurgence as speakers leave the workforce and are no longer subject to the same type of linguistic-marketplace-induced pressure) may be common for non-standard linguistic features. Although relatively little work has yet investigated this claim, Rickford and Price (2013) provide evidence for just this sort of post-adolescent back-off for morphosyntactic features in the speech of African-American females, indicating that standard/nonstandard variant ratios may be a fruitful indicator of age for profiling systems across multiple linguistic levels. Tagliamonte and Baayen (2012) has also demonstrated that non-standard was/were regularization in York English exhibits the type of u-shaped curve associated with post-adolescent back-off and subsequent post-retirement-age resurgence. Additionally, Barbieri (2008) in her analysis of the Longman American Corpus provides evidence that non-standard lexical items (e.g. taboo words, profanity, slang) are significantly more prevalent in the speech of younger speakers (age 15-25) than older speakers (age 35-60).

A small number of studies have also examined age-based differences at the discourse and lexical levels. Stubbe and Holmes (1995) for example found that younger speakers in the Wellington Corpus of New Zealand English favored set marking tags[15] at roughly twice the rate of middle aged speakers, while middle aged speakers tended to use the discourse markers sort of/kind of and I mean/I think at nearly twice the rate of the younger speakers. Likewise, Barbieri (2008) in her investigation of the

[15]I.e. phrases indicating membership in a more general category; e.g. "… and stuff like that," or "…and whatnot."

Longman American Corpus found that speakers in her younger and older age groups favored different subsets of affective markers, stance adverbs, intensifiers, and response tokens, and that the younger speakers (somewhat surprisingly) use politeness markers (e.g. sorry, please) significantly more frequently than the speakers in her older age group. The age-based difference in usage of stance adverbs and intensifiers is particularly pronounced, and has also been noted by Ito and Tagliamonte (2003) and Xiao and Tao (2007). It seems from these investigations that the inclusion of discourse marker, stance adverb, and intensifier distributions may be of some use to ASP systems in predicting age group.

2.3.3 Region

Regional dialect has undoubtedly received the most attention to date of any of the social traits listed in this section. The scope on which one can examine regional dialect runs the gamut from the level of nations (e.g. American vs. British English) to individual social networks distributed across different city-block groupings within the same town (e.g. Milroy, 1980). As the goal of this dissertation is to construct a speaker profiling system for American English, I will focus my discussion here on broad, sub-national geographic regions of the United States. The Atlas of North American English (ANAE; Labov et al., 2006) divides the United States into roughly 10 distinct major dialect areas:

1. West
2. North Central
3. North (subdivided into North and Inland North)
4. Midland
5. South (subdivided into South, Inland South, and Texas South)
6. Western PA
7. Mid-Atlantic
8. New York City
9. Western New England

10. Eastern New England

These dialectal divisions are based primarily on acoustic analysis of the vowels of speakers interviewed as part of the TELSUR project. Roughly, the divisions are based on relative backing, fronting, lowering, raising, and merging of certain vowels indicative of the various sound changes in progress in different regions of the US, as well as glide trajectories and contextual realizations such as the split-/æ/ system. For the purpose of automated speaker profiling however, these geographical distinctions are typically too fine grained to be of use given the amount of data present in spoken-language corpora and the level of metadata available on regional speaker origin. For the purposes of this dissertation, the dialectal divisions of the ANAE will be condensed into a four-way regional classificatory schema (Northeast, South, Midwest, West) corresponding to the U.S. regional mapping used by the Census Bureau. This approach to regional division has the benefits that a) the regions are well defined and do not require a level of speaker metadata below the state level, and b) this regional division schema is comparable to that used by similar existing work in speaker profiling (e.g. Gillick, 2010) and thus enables direct cross-study comparison of results.

It seems likely that phonetic analysis of the key vowels relating to the ANAE dialect regions that correspond to the U.S. Census Bureau regions would prove fruitful in distinguishing speaker region of origin. Such vowel differences are likely the driving discriminatory factor in the high dialect-identification accuracy that Biadsy et al. (2011) achieves with the phone-type supervector approach described above. Interestingly though, none of the work I have been able to find in automated profiling of speaker dialect region to-date attempts to take into account the types of glide trajectory and contextual features detailed in the ANAE. Accounting for such features may result in a significant improvement in regional classification, particularly for speakers from regions primarily distinguished from one another on the basis of these features, such as the

South for glide trajectories and various regions of the Northeast for split-/æ/ systems.

There is unfortunately no comprehensive work detailing consonantal differences between geographical regions of the US, but there exist in the literature a description of a few consonantal features potentially useful for dialect classification. /ɹ/-lessness for example may be of use in distinguishing certain regional dialects along the eastern seaboard (Wolfram and Schilling, 2015). There is also some limited evidence that the degree of closure and voicing in stop consonants may be useful in distinguishing dialect regions with some German substrate influence from other regional varieties (Jacewicz et al., 2009; Purnell et al., 2005).

In addition to phonetic variation there is also of course a great deal of regional lexical variation, and a number of studies have demonstrated the feasibility of mining certain lexical alternations or n-gram features for the prediction of dialect region for author profiling tasks (e.g. Cheng et al., 2010; Eisenstein, 2015; Wieling and Nerbonne, 2010). Gillick (2010) is one of the few studies to predict dialect region using lexical variation in spoken data. He demonstrates that training a Margin Infused Relaxed Algorithm (MIRA) classifier using simple bigram feature vectors extracted from conversational speech resulted in classification of speakers in a four-way dialect categorization task with a relatively high degree of accuracy (56-70% accuracy for Northeast, South, and West regions, 38% accuracy for the Midwest), suggesting that such an approach may prove a useful component in the profiling system proposed here.[16]

Regional variation has also been demonstrated in the morphosyntactic realm. One particularly obvious example of such regional variation is the way in which speakers

[16]While dialectologists tend to focus on semi-salient, specific lexical alternations such as firefly vs. lightning-bug, soda vs. pop, etc., Gillick (2010) and others cited in this paragraph do not place any specific importance on saliency or use alternation lists generated from surveys of dialectology work. Rather, they select n-gram features based on statistical analysis of the data (often via information gain or mutual information ranking). As a result these n-gram features tend to contain more discourse markers, pause fillers, and other frequent lexical items and less of the specific (and typically more rare) content lexemes on which traditional dialectology has focused.

from certain regions produce the 2nd person plural pronoun. Whereas the 2nd plural pronoun 'you' does not differ from the 2nd singular pronoun in 'standard' American English, speakers from the Pittsburgh area are well known for their local variant form 'yins' or 'you-uns' (Johnstone et al., 2006; Johnstone and Kiesling, 2008), speakers of southern varieties tend to favor the elided form 'y'all' (Richardson, 1984; Wales, 2004), and speakers from the mid-west may adopt variable you/yous (Wales, 2004). Likewise the propensity for double modal usage and a-prefixing (as in she's a-comin' home) in southern dialects (though rarely elsewhere) is well documented (e.g. Wolfram and Schilling, 2015, pgs. 378-379). Many such regional-specific syntactic constructions are rare however (with the exception of the 2nd plural pronoun), and so their utility to the project at hand may be limited.

Finally, there is also evidence of regional variation at the discourse level. Tannen's (1984; 2000) work on conversational style for example suggests that speakers from different regions may have different norms when it comes to speaking rate and length of intra- and inter-turn pauses– a finding confirmed by Kendall's (2013) sociophonetic corpus work. It may therefore be useful to include some measurement of speech rate and pause length for regional classification.

2.3.4 Ethnicity

Ethnicity is a particularly tricky trait to handle. First, it should be made abundantly clear that there is scant if any evidence for physiological differences between ethnicities in terms of the vocal apparatus. The only study of which I am aware attempting to scientifically investigate morphological differences in the vocal tract is Xue et al. (2006), who used Acoustic Reflection to measure the oral length, pharyngeal length, total vocal tract length, oral volume, pharyngeal volume, and total vocal tract volume of male

speakers from three different ethnicities in America: African Americans, White Americans, and Chinese Americans. They found no significant differences between African Americans and White Americans in any of the six measurements tested, but found significantly larger oral volume and total vocal tract volume for Chinese Americans as compared to the two other groups. While they suggest this is due to an underlying physiological difference between Chinese Americans and the other two groups, it should also be noted that while African Americans and White Americans spoke American English during their tests, Chinese Americans spoke Mandarin. It's possible that language difference was the driving factor behind their significant results rather than physiological differences tied to ethnicity. I will proceed from the standpoint that any linguistic differences between ethnicities are not a reflection of underlying physiological differences, but are rather due to differences in *ethnolect* or *ethnolinguistic repertoire*.

The term 'ethnolect' has been used in rather different ways over the years. The canonical definition of an ethnolect is a *"[variety] of a language that [marks] speakers as members of ethnic groups who originally used another language or distinctive variety"* (Clyne, 2000, pg.86). Others have problematized this idea of the ethnolect however, pointing out that one does not need to be of a particular ethnicity to speak the ethnolect associated with that ethnicity, and that constructing ethnolects in opposition to the standard can serve to marginalize speakers of that ethnicity (Jaspers, 2008). Some have also argued that talking about groups of dialectal features associated with ethnicities as discrete ethnolects insinuates that such feature groupings are monolithic, ignoring the inherent variation of feature usage that occurs within any sociolect and the interplay that may occur between these features and various higher order meanings within the indexical field. As Eckert puts it:

> *"…not only can the notion of ethnolect serve to reinscribe popular ideologies, it also belies the constructed nature of linguistic varieties and of social (in*

this case ethnic) categories. The term ethnolect (like sociolect and the more generic dialect) reflects a view of language as a fixed rather than fluid entity, and of identity as compartmentalized, allowing one to think of an ethnolect as a discrete system indexical of ethnicity alone."

— Eckert (2008b), pg. 26

For convenience's sake, I will use the term ethnolect in a loose sense to mean those dialectal features which are correlated with a particular ethnicity, bearing in mind of course that these features may index other social meanings, and that not all speakers who identify as members of a particular ethnicity will necessarily use these features to index their ethnicity. I will confine my discussion to literature referring to the four ethnicities which make up the largest percentage of the population of the United States according to the latest census statistics: White, Latino, African American, and Asian (Humes et al., 2011, pg. 4).

Before continuing I would like to note that much of the work on ethnolects sets ethnolectal features in opposition to "standard" American English, and treats white speakers largely as the default "standard" category. There is an underlying assumption that whiteness represents a lack of ethnicity, and that correspondingly any retreat by speakers from their own ethnic varieties represents a corresponding retreat from their ethnicity.[17] This framing of whiteness as the default is unfortunate, and I will attempt to minimize such framing as much as possible in the discussion below.

There is a rich literature examining the group of varieties known collectively as African American (Vernacular) English (AAE). AAE is traditionally characterized by a constellation of non-standard morphosyntactic features such as copula deletion, neg-

[17]White speakers of course also have their own ethnolects, though Anglo varieties are nearly always described in the literature as regional varieties rather than ethnolects (Eckert, 2008b, pg. 27)

ative concord, invariant (habitual) be (and other aspectual markers), absence of the -s morpheme in 3rd person singular, possessive, and plural constructions, and so on (Green, 2002; Labov, 1972a; Rickford, 1999). Various phonetic phenomena have also been linked to AAE, such as consonant cluster reduction, th/dh-fronting, /ɹ/-lessness, and merging of the /ɪ/ and /ɛ/ vowels (Green, 2002; Rickford, 1999). Interestingly, African American speakers also appear generally resistant to vowel shifts occurring among the general (read: "white") populace of the area in which they live (see e.g. Labov et al., 2006). Green (2002) also describes a number of lexical items associated with the use of AAE, though most of these are not exclusive to speakers of AAE. This is of course a rather brief synopsis of features indicative of AAE, and as Yaeger-Dror and Thomas (2010) note, AAE is not monolithic– regional and social variation exist within this variety, and different speakers may vary in terms of exactly which of these features they choose to use and to what extent they deploy them.

Comparatively less has been written about the linguistic features associated with English speakers of Hispanic/Latino descent, and while many of the features characteristic of AAE are common to AAE speakers throughout the United States (albeit at different rates or in different configurations), the generalizability of findings from studies of Hispanic English speakers from one region to another is less well established. Chicano English (CE) spoken in the American southwest is by far the most studied variety of English associated with speakers of Hispanic descent. Fought (2003) describes some distinctive aspects of the phonology of Chicano English, noting a marked tendency for the non-reduction of vowels in unstressed syllables (e.g. [tugɛðəɹ] rather than [təgɛðəɹ]), a general lack of gliding, particularly in the high vowels /i, u/, and a particularly tense realization of /ɪ/. She also notes a variable tendency among the speakers she studies to realize /ɑ/ as /a/, and to front /ʊ/ to /ɨ/. As with AAE (and many other non-standard English varieties), Fought also notes a tendency for CE speakers to realize the interdental fricatives as apical stops, reduce consonant clusters, and to glot-

talize word-final voiceless stops. Likewise Fought finds that CE speakers tend to use a variety of morphosyntactic features common to other non-standard varieties, including negative concord, generalization of was/were for use with plural subjects, and the use of non-standard pronoun forms (e.g. 'hisself'). Interestingly, Fought also notes a few morphosyntactic features which appear particular to CE, including certain patterns of non-standard prepositional and modal usage. Bayley and Santa Ana (2004) also note a number of morphosyntactic features characteristic of Chicano English in the southwest, including absence of the past tense -ed suffix, absence of 3rd singular -s, and the variable absence of direct objects. Prosody may be of particular use in distinguishing speakers of CE (and perhaps other ethnolectal dialects associated with those of Hispanic origin), as Fought (2003) also describes a particularly marked phrase-final rise-and-sustain and rise-and-fall intonation pattern common to this variety.

It is not established exactly how well features described in Chicano English, which is primarily influenced by Mexican Spanish, might generalize to speakers of Hispanic ethnolectal varieties in other parts of the country which have been influenced by other varieties of Spanish. However, emerging research on Hispanic English varieties in the American southeast (primarily around North Carolina) generally agree with the phonetic and morphosyntactic patterns found in Chicano English (see e.g. Callahan-Price, 2013; Kohn, 2008). It seems reasonable therefore to use the detailed descriptions of CE as a jumping-off point for identifying linguistic features which may index Hispanic ethnicity.

So far as I'm aware, no one has undertaken a comprehensive study of Asian ethnolectal varieties as Fought (2003) and Green (2002) have done for Chicano English and African American English, respectively. Some have in fact argued that because of the incredibly diverse ethnic and linguistic backgrounds of those lumped together in the category of 'Asian American' (e.g. Korean-Americans, Chinese-Americans, and so on),

such an endeavor is likely doomed to fail from the start (e.g. Wong and Hall-Lew, 2014). There are however a few studies from which we may draw some distinctive features used to mark Asian ethnic identity. Newman and Wu (2011) for example demonstrate that Chinese- and Korean-American speakers in their study use a significantly 'breathier' voice quality than Latino, African American, and European American participants. Likewise they find that the Asian-American speakers also exhibit longer voice onset time for voiceless stops, and lower realizations of /ɛ/ and /ɪ/ than participants of other ethnicities. Newman and Wu posit that these features, while not necessarily comprising an Asian-American 'ethnolect,' may nonetheless comprise part of an ethnolinguistic repertoire (Benor, 2010) from which these speakers draw to index their ethnic identity. In addition, Bauman (2014) has described a particularly backed and monophthongal /o/ realization in the speech of Asian American sorority members in New Jersey as compared to their white counterparts, which she suggests may function as a marker of Asian ethnic-identity. Interestingly, Kirtley et al. (2016) find a similarly backed and monophthongal /o/ vowel to be distinctive of Hawaii English, a variety spoken in a region in which the majority of speakers are of Asian descent (though they do not specifically compare the speech of different ethnic populations within Hawaii to one another). Finally, Hall-Lew (2009) suggests that Chinese-Americans may lead European Americans in the vocalization of /l/, and Wong (2007) presents evidence that Chinese-Americans in New York resist the short-/ae/ split common to European American varieties of New York City English. While a great deal of inter-speaker variability is noted in most of the studies cited here, these features may nonetheless represent a possible starting point for identifying linguistic features potentially indicative of Asian ethnicity in the dissertation at hand. So far as I'm aware, all studies of ethnolinguistic features associated with Americans of Asian descent have focused on acoustic and phonetic features– there is no evidence of which I'm aware suggesting corresponding morphosyntactic or lexical features associated with Asian ethnicity, other than the propensity for some young

speakers to adopt features of AAE as a means for constructing a non-white identity (e.g. Bucholtz, 2004; Chun, 2001).

The material in this section provides a brief overview of the major findings for ethnolinguistic differentiation among Hispanic, African American, and Asian-American speakers. Before closing this section, it is important to point out again that many of the features associated with the ethnolectal varieties described above are shared among many non-standard dialects across America, such as negative concord and the generalization of is/was to plural subjects. It is therefore not the presence of these individual features per se that should be taken as indicative of a particular ethnolect and by proxy ethnicity, but rather the particular configuration and rates in which they appear.

2.3.5 Education

Of the five social traits of focus for this dissertation, Level of education is by far the least studied and consequently the least well understood in terms of which speech features may be indicative or discriminative. Most of what is known about speech features that may be tied to education comes from linguistic work focusing on education as a foundational aspect of socioeconomic class rather than work focusing on education level itself. Education and SEC are intricately linked, and often co-linear to some degree. Those with higher levels of education also tend to belong to higher socio-economic classes (and vice-versa), and likewise those features which have been found to be associated with higher education levels have also been associated with higher SECs. Often in fact, level of education is one of the key contributing factors in determination of SEC (e.g. Labov, 1966; Trudgill, 1974).

From the sociolinguistic realm, much of the work examining education and SEC has been based on the notion of standard/non-standard feature ratios. Higher education and SEC levels have been repeatedly linked to higher ratios of standard linguistic variants in cases of stable linguistic variation. Labov for instance found that SEC (and consequently education) level clearly delineated speakers with respect to coda-/ɹ/ and th/dh production, with those in the higher classes consistently producing higher standard/non-standard realization ratios across all styles for both variables (1966, pgs. 221-222). Likewise Trudgill (1974) found that speakers in higher socio-economic classes reliably produced higher ratios of standard (velar) to non-standard (alveolar) variants of the ING variable. This production difference appears in some cases to have lead to a higher order indexical association between the standard variant and high levels of SEC and/or education. Campbell-Kibler (2006) for instance has found that listeners in a perception study strongly associated guises using the standard velar variant of ING with being both wealthy and educated. This pattern of social class differentiation among standard/non-standard variants is not confined to phonetics either– Wolfram (1969) for example found a similar pattern in which frequent use of negative concord were correlated with lower social classes in the speech of African American English speakers from Detroit.

Education tends to be notoriously difficult for human forensic speaker profilers to get right. Schilling and Marsters (2015) report that some professional forensic linguists and criminal profilers with whom they are in contact have noted that it is not uncommon for a speaker or author to have great facility with "educated language" despite having relatively little formal education, and have consequently suggested that attempting to predict education level should be done with extreme caution, if at all. While this speaks to the potential difficulty of estimating education level in an automated speaker context, it also makes it an interesting and potentially worthwhile subject of exploration. Unfortunately, as Nguyen et al. (2013) note, very little attention has

been paid so far to the investigation of education level or SEC within the realm of computational (socio)linguistics. As far as I'm aware, the only study so far which has attempted to predict education level from American English spoken language data is Gillick (2010), who achieved 67% unweighted accuracy in identifying education level operationalized as a four-way binned classification problem. However, some recent work within sociophonetics has demonstrated that level of education (and indeed even the degree to which the educational institutes attended are locally vs. nationally oriented) may have a demonstrable effect on the degree to which speakers participate in local dialectal phenomena. See e.g. work by Prichard and Tamminga (2012) and Fisher et al. (2015) for a discussion of such effects of education on realizations of the Philadelphia short-a system. This speaks to the importance of including education as a jointly modeled social strait when designing a speaker profiling system aimed at predicting region (and likely other social traits), and hints that features related to vowel production and other local dialectal phenomena may hold some power in education prediction tasks.

2.4 Computational Foundations

In this section I discuss a few of the key computational and machine learning concepts used in the automated speaker profiling models discussed in chapters 5 and 6. I begin with a brief overview of neural network architecture in general. Following this, I discuss the operationalization of a multi-task learning framework within the context of neural networks, and detail some of the benefits that multi-task learning can provide when dealing with multiple related prediction tasks such as those integral to this dissertation.

2.4.1 Neural networks overview

An in depth discussion of all possible types and parameters of neural networks is beyond the scope of this section. I will instead focus here on a bird's eye view of the basic components and functions underlying the most common neural network designs, with an eye to making the technical discussions in the following chapters more accessible to less technically-inclined audiences.

All neural nets are at heart made up of a number of interconnected yet individual processing nodes (i.e. 'neurons'). While each individual node is relatively uncomplicated, functioning in unison these groupings or 'nets' of nodes can be harnessed to solve highly advanced classification problems which cause difficulty to traditional algorithmic approaches. I'll start by discussing the basic anatomy of an individual node, then proceed to a discussion of how these nodes are organized into groupings, or 'layers', designed to perform specific functions within the net at large. The basic architecture of a neuron is best exemplified by explaining the inner workings of a single-input neuron. A diagram of such a neuron is presented in figure 2.1 (reproduced from Hagan and Demuth, 1995, pg. 38).

The output of a neuron a is determined by the application of an activation function[18] $f()$ to the sum of the product of the scalar weight w and the scalar input p plus the bias b (in figure 2.1 the bias is 1). In mathematical notation therefore, the output of a single input neuron can be represented by the equation $a = f(wp+b)$. The output of a neuron with an input of 4, a weight of 5 and a bias of 1 therefore would be calculated as $a = f(4 * 5 + 1)$, or $f(21)$. The actual output a depends upon the type of activation function used. Activation functions come in many different flavors (popular activation functions include the sigmoid, tanh, and ReLU functions), and the

[18]Sometimes the activation function is referred to as a 'transfer function'

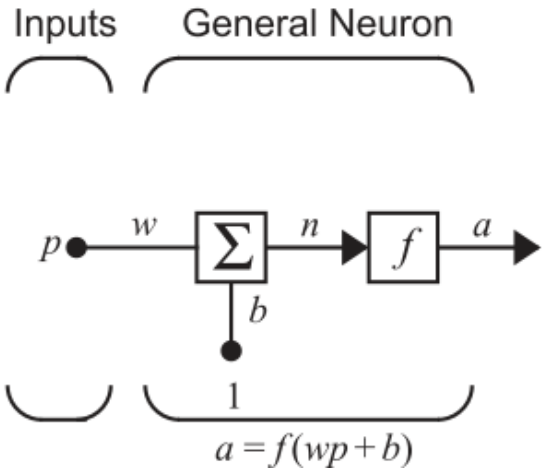

Figure 2.1: Single input neuron

choice of activation function will depend on the type of problem at hand. The weight roughly corresponds to the 'synaptic strength' of a biological neuron, and modulates the threshold at which the neuron will become activated. Both w and b are typically adjusted by a learning rule such that the input/output relationship of the neuron meets a specific goal. In practical applications, it is more common to use multiple-input neurons. Multiple-input neurons function much the same as single-input neurons, except that instead of a scalar the input is now a vector of length R (R = number of scalar inputs), and instead of a scalar weight w a weight matrix W composed of R columns is used.

It is uncommon for a single neuron to be sufficient for real-world tasks. Often groups of neurons operating in parallel are used instead. These groups are called 'layers.' A diagram of a typical neuron layer is presented in figure 2.2 (reproduced from Hagan and Demuth, 1995, pg. 44). In a neuron layer, each element p_i of the input vector is connected to each neuron through the weight matrix W (which now has S rows, S corresponding to the number of neurons in the layer. Each neuron uses its corresponding row S_i in the weight matrix). Each neuron has its own bias element b_i within the bias

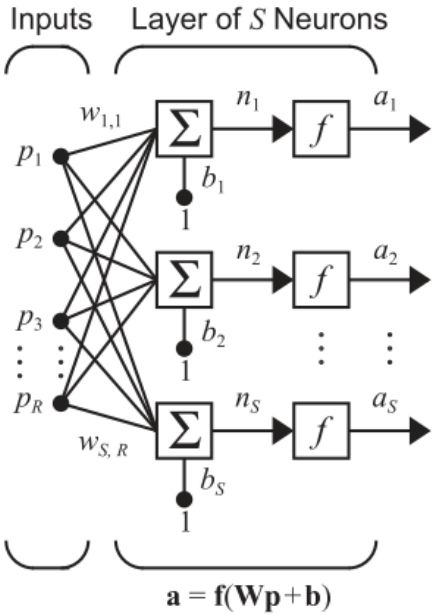

Figure 2.2: Multiple input neuron

vector and its own output a_i. It is not uncommon for a neural net to be composed of multiple layers, each successive layer taking as its inputs the outputs of the previous layer, depending on the task at hand. Multi-layer networks are particularly powerful. Whereas single-layer networks typically are only capable of classifying linearly separable patterns, multi-layer networks may be used for arbitrary classification problems and can serve as universal function approximators (Hagan and Demuth, 1995, pg. 397). In multi-layer nets, the layer whose output is also the network output (i.e. the final layer) is termed the 'output layer,' and layers in between the output layer and the input layer are called 'hidden layers.' For classification tasks, if the desired output is binary (1 or 0), a single neuron using a binary activation function can be used in the output layer. In multi-class prediction tasks, often the output layer will consist of a number of neurons corresponding to the number of possible classes, each using some sort of sigmoid activation function. In these cases the outputs of each neuron in the output layer may be interpreted as class probabilities.

Training a neural network involves presenting the net with multiple examples of input data and updating the weight and bias matrices according to a learning rule (also called a 'training algorithm'). Learning rules fall into 3 general categories: supervised, reinforcement, and unsupervised. For supervised learning, the net is presented with a set of input examples and desired outcomes, or 'target outputs.' Supervised training algorithms compare net outputs to target outputs, and update the weight and bias matrices accordingly. Reinforcement learning is much like supervised learning, except that rather than being provided with specific target outputs, the net outputs are instead given a grade or score according to how well they match desired network performance. In unsupervised learning the training algorithm is not provided with any target output or grade/score, but instead updates the weight and bias matrices solely on the basis of network inputs. Unsupervised algorithms all use some type of clustering operation to categorize the input into a finite number of classes.

For supervised neural nets of the type that are used in this dissertation, the standard training algorithm used to update the weight matrix is called 'gradient descent.' Briefly, gradient descent is an optimization function designed to iteratively tweak the parameters of a model in order to minimize an error function.[19] First the derivative (or 'gradient') of the error function is calculated using the network output based on the current model parameters. This error gradient is then propagated backwards through each successive layer of the network in order to determine the gradient of the error function with respect to each connection weight in the weight matrix, essentially assigning a proportional degree of 'blame' to each neuron in the network (propagation of the error gradient backwards through the network is referred to as back-propagation, or backprop for short). Once the gradient of the error function is computed with respect to each connection weight, the weight matrix W is updated in the opposite direction of

[19]The error function estimates the degree of error between current model output and target model output. Which error function one uses will be specific to network design and data

the error gradient $\nabla_W Error(W)$ in a step proportional to the learning rate η:

$$W^{t+1} = W^t - \eta \cdot \nabla_{W^t} Error(W^t)$$

The learning rate controls the speed with which the network converges on an error minimum, and can have detrimental effects if improperly calibrated. Too large a learning rate will cause the algorithm to overshoot the error minimum and diverge. Too small a learning rate will necessitate an excessive number of training iterations before converging.

Network training takes place over several successive cycles through the training set, or 'epochs.' How often the weight matrix is updated during an epoch depends on which flavor of gradient descent one uses. Networks trained using stochastic gradient descent (SGD) perform a weight matrix update after every training example encountered. On the other end of the spectrum is batch gradient descent, wherein a weight matrix update is performed only once at the end of each epoch. Between these two extremes is mini-batch gradient descent, wherein the weight matrix is updated once at the end of every batch of m training examples encountered. Batch gradient descent will gently minimize the error function until reaching a local minimum, but is relatively slow to train since weight updates are only performed once every epoch. Stochastic gradient descent will approximate a local minimum much more quickly, but is inherently unstable due to the frequency of weight updates. SGD has a tendency to bounce around a local minimum instead of finding the center (while this leads to poor convergence on local minima, a certain level of instability can actually be a good thing, as it provides a mechanism for escaping, or 'bouncing out of,' sub-optimal local minima). Mini-batch is a good compromise between the speed of SGD and the stability of batch gradient descent, and is often the flavor of choice when training neural networks.

Often, the surface of the error function in problems for which a neural network approach is beneficial is non-convex, meaning the error surface in addition to a global minimum will include saddle points and sub-optimal local minima which can trap a network trained with vanilla gradient descent. A common method of escaping saddle points and sub-optimal local minima is to include a momentum term while performing the gradient descent weight update. The momentum term is a fraction γ of the previous update gradient μ:

$$\mu^t = \gamma \cdot \mu^{t-1} + \eta \cdot \nabla_W Error(W^t)$$

$$W^{t+1} = W^t - \mu^t$$

The addition of the momentum term accelerates movement along the error surface when the error gradients at training step t and training step $t-1$ point in the same direction, and decelerates movement along the error surface when they point in different directions. This helps to dampen oscillation along the walls of saddle points, and, akin to a ball picking up speed as it rolls down a hill, can help the algorithm power through plateaus and small dips in the error surface (i.e. sub-optimal local minima) when it encounters them. Various modifications to the standard momentum term (e.g. Nesterov Acceleration) and optimizations of the gradient descent algorithm (e.g. AdaGrad, RMSProp, Adam, etc.) aimed at escaping sub-optimal local minima and decreasing training time are commonly used when training neural networks, but an in depth discussion of all of these is unwarranted here.

Finally it should be noted that while neural networks can be extremely powerful tools, they also have a propensity to overfit the training data if improperly designed. Neural nets which have been overfitted have essentially 'memorized' the training data

rather than learning the general patterns within it, resulting in poor generalizability and consequently poor performance on unseen data of the same kind. There are several techniques used in practice to reduce the probability of overfitting a neural net during training.

Often, overfitting is caused by using a network design that is more complex than the problem or data warrant. The complexity of a given net is dependent on the number of adjustable 'free' parameters (weights and biases) contained within it, which is of course in turn dependent on the number of neurons and neuron layers included within the net. A net with too many free parameters, in other words too much flexibility, simply devolves into a look-up table. 'Growing' and 'Pruning' are two strategies that seek to minimize the probability of over-fitting by constraining the number of neurons present in the net, thereby constraining net complexity. 'Growing' refers to methods in which one starts with zero neurons and then successively adds neurons one-at-a-time or in blocks until network performance reaches some sort of desired threshold. 'Pruning' is roughly the opposite, in which one starts with a large number of neurons (with consequently a high probability of overfitting), and then successively removes them until network performance degrades significantly.

Aside from adjusting the number of neurons present in the net, there exist a number of common regularization strategies aimed at minimizing the chance of overfitting by modifying the training strategy. Perhaps the most popular of these regularization techniques is 'dropout.' Dropout refers to the process by which, at every training step, each neuron in the net (with the exception of the output neurons) is assigned a probability p of being entirely ignored, or 'dropped out' for that particular training step. This prevents neurons from co-adapting with neighboring neurons or relying too intently on only a few input neurons. As such, it causes the neurons to be less sensitive to small variations in the input, and consequently makes the net as a whole more robust and

less prone to over-fitting. Another common regularization strategy that modifies the training process is 'early stopping.' This technique takes advantage of the fact that when a network begins to over-fit the training data, network performance on the validation set begins to degrade. When applying early stopping, the network is tested against a validation set every x number of training steps. Every time performance on the validation set reaches a new peak, a snapshot of the network in its current state is saved. Once performance on the validation set hasn't peaked for a certain number of training intervals, training is halted and the last 'peak' snapshot of the model is restored, effectively halting network training prior to the point at which overfitting began to occur.

In addition to modifying the network or the training process itself, regularization may also be achieved via data augmentation. Data augmentation refers to the generation of new training data from existing training data, modifying the new data slightly in a variety of aspects to which one desires the network to be robust. For example, in training neural networks for image classification, one could generate new training data by slightly shifting, resizing, rotating, or blurring existing data. Such augmentation makes the network more robust against these type of operations in unseen data by minimizing the probability that the model will overfit the peculiarities of the training data in these regards.

Though many more regularization strategies may be included depending on network design, data, and so on (e.g. max-norm regularization, ℓ_1 and ℓ_2 regularization, etc.), those techniques detailed above are some of the most effective and consequently widespread.

Having laid out the basic structure and function of neural networks along with a few of their pitfalls and corresponding remedies, I now turn to a more pointed discussion

of the type of network architecture I plan to use in this dissertation: Multi-Task neural networks.

2.4.2 Multi-Task Learning (MTL) and joint-prediction

Multi-Task Learning (MTL) within a Machine Learning context refers to the process of constructing a model such that it learns to perform several related tasks at the same time while using a shared representation. The basic idea is that by using a shared representation, what is learned for each task can help other tasks be learned better– in other words, training signals for all tasks serve as an inductive bias for each task learned in parallel (Caruna, 1997, pg. 41). One of the more prevalent approaches to MTL in general is its extension to support vector machines, introduced by Evgeniou and Pontil (2004). However, as Parameswaran and Weinberger (2010) point out, MTL operationalization within an SVM requires each of the different learning tasks to share the same set of classes, which makes such an approach unsuitable for many NLP-related applications, as well as this dissertation. NLP researchers instead have tended recently to rely on deep neural networks to operationalize an MTL approach. MTL operationalized within neural networks has been used to great effect in a number of NLP-related problems, including semantic classification and information retrieval (Liu et al., 2015), Multiple language translation (Dong et al., 2015), and joint POS tagging, chunking, NER, and semantic role labeling (Collobert and Weston, 2008).

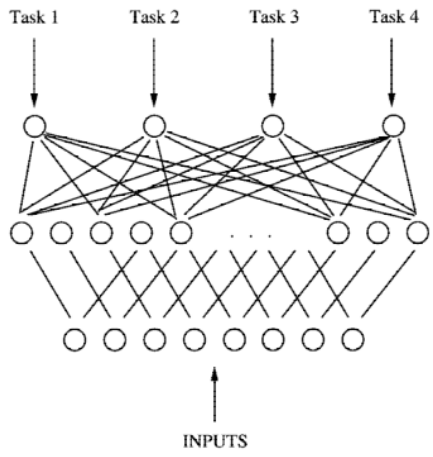

Figure 2.3: Multi-task net

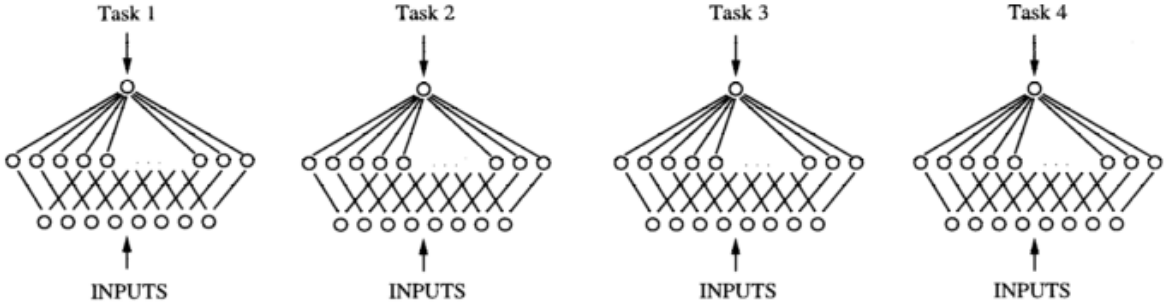

Figure 2.4: Single-task net

In the context of neural networks, whereas a traditional STL approach to training learners on four related tasks would require four separate nets, each with an individual output related to the task at hand, MTL is operationalized by combining these individual nets into a single net with four separate output layers. Each output layer corresponds to one of the four individual tasks, and each of these tasks share the hidden layer(s) and the inputs (the hidden layer(s) here functions as the 'shared representation' necessary for MTL). A graphical representation of an MTL neural net architecture is provided in figure 2.3 (reproduced from Caruna 1997, pg. 44). Compare this to figure 2.4 (reproduced from Caruna 1997, pg. 43), which shows how these same four tasks would be instantiated in an STL framework.

MTL is a particularly attractive concept for the purposes of this dissertation in that it speaks in part to the principle of intersectionality described in section 2.2.3. As Caruna notes, MTL:

> *…allows features developed in the hidden layer for one task to be used by other tasks. It also allows features to be developed to support several tasks that would not have been developed in any STL [Single Task Learning] net trained on the tasks in isolation. Importantly, MTL also allows some hidden units to become specialized for just one or a few tasks; other tasks can ignore hidden units they do not find useful by keeping the weights connected to them small*
>
> — Caruna (1997), pg. 26

In other words, such a system applied to this dissertation would be able to learn complex relationships between features related to the predication of various social trait classes (i.e. classification tasks), and develop co-constructed features within the hidden layer(s) that are jointly informative for multiple traits or trait groupings.

In addition to improving generalized performance, MTL also improves computational efficiency. As Caruna (1997) notes, training an MTL net on a set of tasks often requires less computation than training the individual STL nets which would be required for learning each task individually.

Chapter 3: Data

This chapter provides a comprehensive overview of the corpus, social trait operationalization and data featurization procedures used in this dissertation.

3.1 Corpus

The primary data source for this dissertation is the 2008 National Institute of Standards and Technology Speaker Recognition Evaluation corpus (2008 NIST SRE). In total, the training sets[1] [2] and test set[3] comprise roughly 2,500 hours of multilingual telephone and interview speech gathered throughout 2007, divided into subsets based on recording length (10 sec, 3 min, 5 min, 8 min, 12 min), and recording type (interview, phone call). Each sound file in the NIST SRE corpus is accompanied by a time aligned transcript generated by an automatic speech recognition system deployed by the creators of the corpus.[4]

[1] training set part 1: https://catalog.ldc.upenn.edu/LDC2011S05
[2] training set part 2: https://catalog.ldc.upenn.edu/LDC2011S07
[3] test set: https://catalog.ldc.upenn.edu/LDC2011S08
[4] Exactly which ASR system was used to generate the transcripts is not made clear in any of the NIST SRE documentation.

3.1.1 Corpus metadata

The following metadata is available for most speakers in the NIST SRE corpus:

- Gender
- Birth Year
- Ethnicity
- Years of Education
- Occupation
- Native Language
- Age English Acquired as 2nd Language
- Other Languages Spoken
- Country Born in
- Country Raised in
- State Born in
- State Raised in
- City Born in
- City Raised in
- Smoker
- Height
- Weight

In addition to the speaker metadata, the following metadata is available for most recordings:

- Recording Date
- Conversation quality (good, acceptable, NA)
- Signal quality (good, acceptable, NA)
- Call language
- Microphone type
- Telephone type

3.1.2 Corpus subsets used here

Recordings in the interview subsets of the corpus are (surprisingly) generally of lower recording quality than those in the telephone subsets, and typically were recorded using one microphone to capture speech from both participants in the interview. As

a result, the automatically generated transcripts for interview speech are more error prone, and do not indicate which speaker is speaking at any given time. For this reason, this dissertation focuses on the telephone subsets of the NIST SRE corpus. The telephone subsets are separated into 10 second subsets and 5 minute subsets, each recording containing a separate channel for each interlocutor. As 10 seconds is likely not enough speech to exemplify many of the non-acoustic features on which the ASP models presented in this dissertation rely, the data used here come exclusively from the 5 minute telephone subsets of the NIST SRE corpus (the two relevant subsets are termed "short2" and "short3" in the corpus nomenclature).

As the purpose of this dissertation is to construct ASP models focused on speakers of American English, recordings of speakers who were raised outside of the US are excluded, as are recordings of speakers who were raised in the US but who did not acquire English as a second language until after the age of 5, and/or who do not consider themselves to be native speakers of American English. To give the ASP models the best chance of success, recordings with a conversation or signal quality rated as anything less than "good" are also excluded, as are recordings made using a telephone speaker-phone. In most cases both channels from a given telephone conversation in the corpus meet this criteria. There are however instances where only one of the two recorded channels from a given conversation met these criteria and thus only one side of that conversation (i.e. speech from only one of the two interlocutors from that conversation) is included in the present analysis.

Applying these filtering criteria results in 1,001 remaining 5 minute single-channel recordings (approximately 83.4 hours of recorded speech in total) from 669 unique speakers. A break down of speakers with respect to all five social traits examined in this dissertation is available in chapter 4.

3.1.3 Transcript accuracy

As a great deal of the feature extraction and analysis depends on the accuracy of the time aligned transcripts included in the corpus for each sound file, it is important to perform an assessment of said accuracy. The standard measure of accuracy applied to automated speech recognition systems is word error rate (WER), defined as the sum of the number of substitutions, insertions, and deletions of lexemes required to match the ASR output to the reference, divided by the true number of lexical items in the reference:

$$WER = \frac{S + D + I}{N}$$

Though human performance on transcription tasks is often cited as having around a 4% word error rate (see e.g. Lippmann, 1997), Xiong et al. (2017) report that professional human transcribers hired to transcribe conversational telephone speech from the Switchboard and CallHome sections of the 2000 NIST eval test set produced transcripts with word error rates of 5.9% and 11.3% on average, respectively.[5] As the Switchboard and CallHome data from the 2000 NIST eval corpus is quite similar in nature to the conversational telephone speech from the 2008 NIST SRE corpus used in this dissertation, it is reasonable to expect a similar word error rate were we to hire our own professional transcribers to produce transcripts. So long as the automatically generated transcripts in the NIST SRE Corpus exhibit word error rates relatively close to the 6-11% WER reported by Xiong et al. (2017) for human transcriptions of telephone conversations, they should be considered to be acceptable for use within the present dissertation.

[5]The ASR system detailed by Xiong et al. (2017), one of the most accurate ASR systems proposed to date, performed at a WER of 5.8% and 11.0% on the Switchboard and CallHome data– narrowly beating human performance.

To assess the accuracy of the transcripts included in the telephone conversation sections of the 2008 NIST SRE corpus, 8 sound files balanced for speaker sex and ethnicity[6] were selected at random from the subset of NIST SRE telephone conversations used in the present dissertation and transcribed by hand by a professional linguist with extensive experience in transcribing recorded conversational speech (i.e. me, the author of this dissertation). Word Error Rates were then calculated between the automatically generated transcripts and the human generated transcripts. Results of this investigation are presented in table 3.1. It should be noted that the guidelines followed for WER calculation were quite strict, penalizing the automatic transcriptions for lexeme guesses that were phonetically identical yet not lexically identical to what the human transcriber considered to be the true reference lexemes (e.g. "too" instead of "to" was penalized as one substitution, "a just" instead of "adjust" was penalized as one substitution and one insertion, and so on). As such, the WERs presented in table 3.1 should be considered as an accurate assessment of the lexical inaccuracy of the transcripts, but a slightly inflated assessment of the phonemic inaccuracy of the transcripts.

Table 3.1: Word error rate for NIST SRE auto-generated transcripts

	sex		
ethnicity	f	m	Average
Hispanic/Latino	16.5%	10.4%	13.5%
Asian	8.5%	13.9%	11.2%
African-American	13.6%	16.5%	15.1%
White	7.9%	7.6%	7.8%
Average	11.6%	12.1%	11.9%

The observed word error rates of between 7.6% and 16.5% (average: 11.9%) are surprisingly good, and should be well within the bounds of acceptability for the present

[6]Though it would be preferable to examine a selection of sound files balanced for all five social traits examined, this would necessitate a human review of at least 480 individual sound files (roughly 40 hours of speech recordings), which is unfeasible for the present analysis given time and resource constraints.

use case. However, though the sample size is small, these results do appear to hint at ethnicity effects on automatic transcription accuracy– a phenomenon that Tatman and Kasten (2017) have found in several publicly available ASR systems. Though it is possible that higher rates of transcription inaccuracies for African-American and Hispanic speakers may pose problems for syntactic parsing and phonemic alignment/summarization components downstream, the WER does appear to still be within reasonable bounds for these speakers. Such differences in transcription accuracy, while undesirable from the standpoint of ASR systems deployed in the real world, may actually be a source of useful training signals in differentiating speakers of different ethnicities for ASP systems if a measure of discourse coherence and/or disfluency is included.

3.2 Data Preprocessing

Data files for each 5 minute telephone conversation in the 2008 NIST SRE corpus come in the form of a dual channel .sph sound file (one speaker per channel) along with an associated .cfm transcript file of the conversation, marked for current speaker, and time aligned at the word level.

In order to extract exemplars of specific phones from the sound files during construction of the predictive phonetic features outlined in below, the existing lexically aligned transcript must be augmented by a transcription of the sound file aligned at the phonemic level. Several off-the-shelf software tools are available for automatic phonemic alignment, provided one already has a transcript which is time aligned at the word or utterance level, including the Forced Alignment and Vowel Extraction (FAVE) program suite (Rosenfelder et al., 2011), Prosodylab Aligner (Gorman et al., 2011), and the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). Of the forced alignment tools publicly available, MFA is by far the best documented and most up-to-date

(FAVE for instance appears to be no longer actively maintained). MFA also produced by far the best and most consistent phonemic alignment of all the tools tested with sample NIST SRE data.[7] For these reasons, MFA is used as the forced aligner of choice to phonemically align transcripts in this dissertation. The US English version of the Carnegie Mellon Pronouncing Dictionary (CMUDict)[8] was used as the underlying pronunciation reference during forced alignment.[9] All tokens which were out of vocabulary for the CMUDict were transcribed by the author according to the proper format and added to the working lexicon.

MFA requires sound files in .wav format and transcripts in the form of Praat textgrids with utterance annotations of roughly 30 seconds or less. Therefore, prior to feeding data to MFA, the dual channel .sph files are separated by channel in order to produce one sound file per speaker and converted into .wav format. The .cfm transcripts are also separated by channel and converted to textgrid format in order to produce one textgrid per speaker, time aligned at the word level. Because the time alignment provided by the ASR transcripts often fudges the word boundaries by a tenth of a second or so, words which the ASR transcripts indicate as having less than a 0.5 second pause between them are grouped into contiguous utterances and padded with 0.015 second buffers on either side of the utterance in order to let the forced aligner decide where the word boundaries are within the utterance. In other words, the time alignment given in the ASR transcripts is used only to determine utterance boundaries within the sound file– lexeme and phoneme boundaries are determined downstream by the forced aligner. These formatted sound files and textgrids are then passed to MFA, which

[7]Based on visual inspection of the resulting textgrids.

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

[9]It should perhaps be noted that in cases where pronunciation within the speech file doesn't match the pronunciation reference for a given token in CMUDict, phonetic transcription will be inaccurate. For example, if a speaker uses the non-standard alveolar nasal in word-final position for a token such as "running" yet CMUDict does not include a pronunciation using the the alveolar nasal, the word final nasal phone will be erroneously transcribed as the velar variant. Some level of this type of inaccuracy is unavoidable without a major review and revision of the CMUDict, which is beyond the scope of this investigation.

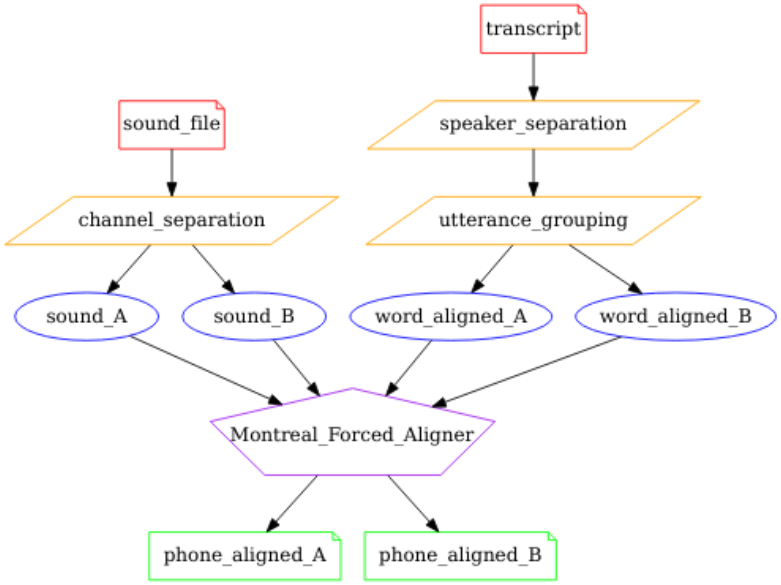

Figure 3.1: Preprocessing pipeline for sound files and transcripts

outputs a textgrid time aligned at the lexeme and phoneme levels for each speaker. The preprocessing pipeline is diagrammed in figure 3.1.

A snapshot of a time-aligned textgrid resulting from passing sample NIST SRE data through the preprocessing pipeline is provided in figure 3.2. Please note the surprisingly good (for telephone data) quality of the sound file, exemplified by the clarity with which formants are depicted in the spectrogram, and the accuracy of the phonemic alignment.

3.2.1 Utterance chunking

Though the NIST SRE corpus is one of the largest compilations of conversational spoken data publicly available, treating each of the 1,001 five-minute recordings that result from the filtering process described above as monolithic (i.e. one data point per recording–

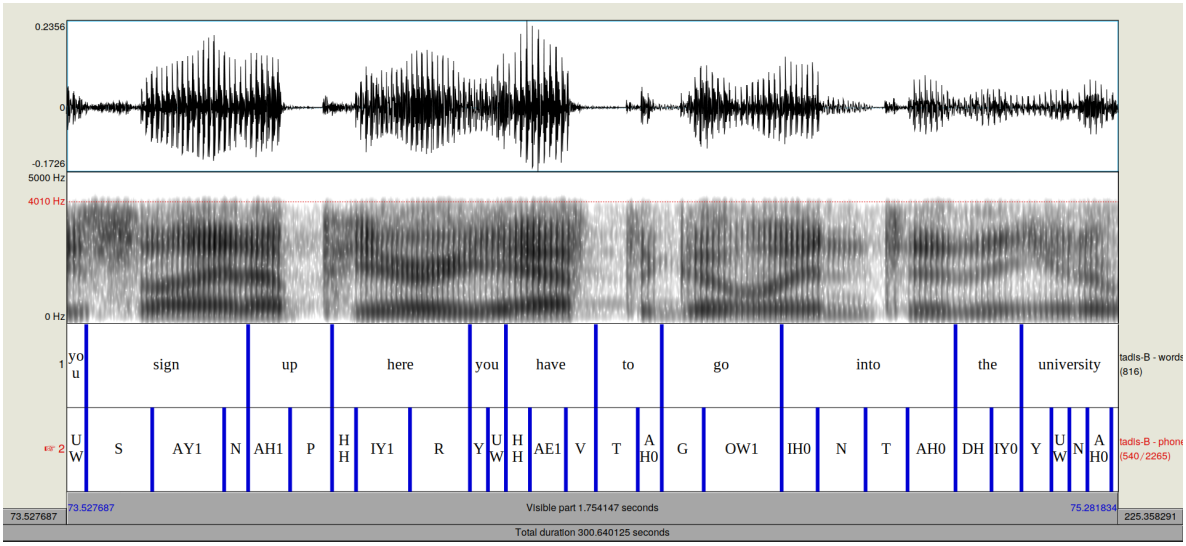

Figure 3.2: Aligned textgrid output sample from MFA on NIST SRE data

1,001 data points) results in too few data points to be of use to neural network models of the type used in this dissertation. For this reason, each of the time aligned transcripts were subsequently chunked into 10 roughly 60 second segments. These segments were produced by randomly selecting contiguous utterance groups (e.g. groups of lexemes with less than 0.5 seconds of silence between them) from each transcript and concatenating them until the resulting segment reached a duration of at least 60 seconds.[10] Performing this chunking operation transforms the corpus from 1,001 five minute speech recordings to 10,010 60 second speech segments. These 10,010 60 second speech segments are treated as the atomic unit of analysis throughout this dissertation. While still somewhat on the low side for typical neural network applications, 10,010 data points should be sufficient to train and test the neural network models presented in chapters 5 and 6.

To sum up, the preprocessing steps outlined here result in 10,010 phone- and lexeme-aligned transcripts, each consisting of approximately 60 seconds of recorded

[10]Though this process necessarily means that some utterance groups from the original recording are present in multiple resulting segmental chunks, this should not be detrimental to the feature extraction or analysis procedures outlined in the following chapters.

speech randomly sampled from a single individual during a single conversation, drawn from 669 unique speakers.

3.3 Social Trait Operationalization

3.3.1 Sex

Speaker sex is operationalized as a two-way classification problem with reference categories corresponding to the self-identification provided by the NIST SRE corpus participants (male/female). All speech segments used in this dissertation came from speakers who self-reported their sex within the NIST SRE corpus metadata, and thus all 10,010 speech segments were included in all sex-prediction tasks.

Though the category as listed in the NIST corpus metadata is labeled "sex" rather than "gender," it is likely that participants in the NIST SRE corpus interpreted this part of the metadata survey as referring to their gender identity, and unlikely that corpus administrators cross-referenced hormone levels, birth certificates, etc. in order to verify that reported "sex" coincided with some definition of biological sex. As such, it may be more appropriate to term this category as "gender" rather than "sex." However, as the exact nature in which this data was collected is not made clear in the corpus metadata, I will stick to the terminology used in the metadata and refer to this category as "sex" throughout this dissertation.

3.3.2 Ethnicity

In order to ensure sufficient data points for each ethnicity considered in the ethnicity prediction tasks detailed in the following chapters, only ethnicities with which at

least 5% of the participants in the subset of the NIST SRE corpus used in this dissertation self-identified were considered (four ethnicities fit this criteria in the corpus: Hispanic, Asian, Black, and White). Of the 10,010 speech segments used as data points for training and testing purposes, 8,660 were drawn from speakers self-identifying as one of these four ethnic categories. Ethnicity was considered as "missing" for the remaining 1,350 speech segments and thus these segments were effectively ignored for all ethnicity prediction tasks. Ethnicity is therefore operationalized in this dissertation as a four-way classification problem with reference categories corresponding to ethnicity self identification provided by the corpus participants (Hispanic/Asian/Black/White).

It should perhaps be noted that the original labels for these four categories in the NIST SRE 2008 corpus are actually "Hispanic/Latino", "Asian", "Black/African American", and "White". The category labels "Hispanic/Latino" and "Black/African American" have been shortened throughout this dissertation to "Hispanic" and "Black" respectively. Noting the "and or" nature of these two categories as defined in the corpus is particularly of note in the case of speakers identifying themselves as Hispanic/Latino, as these two terms don't necessarily coincide. Some fuzziness in terms of what this category might have meant to participants may in part be responsible for the difficulty of accurately identifying speakers belonging to this category that is reported in chapters 5 and 6. That said, none of the ethnic categories as defined in the NIST SRE corpus are necessarily monolithic, and all may encompass distinct micro-ethnic identities which speakers may see at some level as being in opposition. Nor is it necessarily the case that participants internally identified with one and only one of the presented categories, though they were constrained by corpus administrators into choosing only one.

3.3.3 Age

Numerical age for each speaker was calculated based on a speaker's self-reported birth year and the year of corpus collection (2007). A speaker reporting in 2007 having been born in 1989 was therefore treated as being 18 years old (2007 - 1989 = 18). Those with birth years recorded in the NIST SRE metadata which would result in an improbably young or old age were excluded from age prediction tasks, their birth year being treated as "missing." The cutoff for age improbability was a calculated numerical age less than 0 or more than 100.[11] Of the 10,010 speech segments used as data points in this dissertation, 9,660 were drawn from speakers with viable reported birth years (i.e. not missing or clearly erroneous).

For the purposes of this dissertation age is treated as a categorical variable, binned into five distinct categories: 16-25, 26-35, 36-45, 46-55, and 56+. Category boundaries were chosen based on the distribution of numerical age in the corpus[12] as well as with an eye to roughly capturing certain facets of life-stage on either end of the spectrum (i.e. 16-25 likely captures most participants currently undergoing some type of university or continuing education, 56+ likely captures most participants who have retired). While not precisely scientific, there is no real consensus within the relevant literature as to how to bin age effectively for linguistic experimentation and thus no "best practice" to follow in this regard. The category boundaries used here result in a number of bins and bin size that roughly coincides with those used in most existing work on automated age prediction (e.g. Gillick, 2010).

[11] It should perhaps be noted that no speaker in the corpus reported a birth year that would result in a numerical age between 0 and 18 or between 90 and 100. There were however clearly erroneously recorded birth years of 3000, 1850, and so on.

[12] E.g. 56 was chosen as the cutoff age for the oldest age group as data from speakers above this age is sparse and further group delineation would not result in sufficient data points per bucket.

3.3.4 Region

Region is operationalized as a four-way classification task with reference categories corresponding to the four macro-region divisions used in regional mapping by the U.S. Census Bureau (West, Midwest, South, Northeast). This four-way categorization schema was chosen rather than a more sophisticated regional classification system such as that given by Labov et al. (2006) for instance because of A) metadata availability (most participants in the corpus subset used here had viable information entered for "state raised in" but not necessarily for "city raised in", and thus a classification schema operationalized at the state level was preferred) and B) the relatively small size of the corpus subset used (a classification schema with numerous regional categories would result in too many extremely under-represented categories and thus be unsuitable for use with the types of models used here).

Participants were assigned one of these four regions based on the state they reported having been raised in and that state's corresponding region within the Census Bureau regional schema. If a participant listed more than one state in this metadata field, their region of origin was considered "missing" and they were excluded from regional prediction tasks. Of the 10,010 speech segments used in this dissertation, 9,010 were drawn from speakers who could be assigned a region according to this classification schema. The remaining 1,000 segments were excluded from all regional prediction tasks.

3.3.5 Education

Education is operationalized as a three-way classification task based on the number of years of education self-reported by NIST SRE corpus participants. Category boundaries

are chosen so as to correspond with the major delineations present in the American Educational system (primary education, undergraduate education, graduate education). Participants reporting between 0 and 12 years of education are classified as "no-college," those reporting between 13 and 16 years of education are classified as "college", and those reporting 17+ years of education are classified as "post-college". Of the 10,010 speech segments used in this dissertation, 9,220 were drawn from speakers who self-reported the number of education years they had undergone at that time and thus were able to be assigned one of these three categories. The remaining 790 speech segments were excluded from all education prediction tasks.

3.4 Feature Extraction

From each of the 10,010 speech segments used in this dissertation, a host of acoustic, phonetic, and lexical features were extracted and concatenated into a feature vector. These 10,010 feature vectors constitute the data points used in the training and testing of the models presented in chapters 5 and 6. Each individual feature extracted is listed under its corresponding category below, along with a description and notes on operationalization where appropriate.

3.4.1 Acoustic features

Harmonic to Noise Ratio (HNR): HNR is operationalized as the mean harmonic to noise ratio (also called "harmonicity" or "acoustic periodicity") across all vowel chunks for a given speech segment. HNR was extracted using the harmonicity extraction functions available within Praat.

Pitch: Three measures of pitch are taken from each speech segment: mean pitch, max

pitch, and min pitch. These correspond to the average, maximum, and minimum values in Hz for F0 over all vowel segments present within the span of a given speech segment. F0 measurements were extracted using Praat.

Jitter: Three measures of Jitter are taken from each speech segment: Local (the average absolute difference between consecutive periods, divided by the average period), RAP (Relative Average Perturbation: the average absolute difference between a period and the average of it and its two neighbors, divided by the average period), and PPQ5 (five-point Period Perturbation Quotient: the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period). Jitter measurements were extracted using Praat.

Shimmer: Two measures of Shimmer are taken from each speech segment: Local (the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude), and APQ3 (three-point Amplitude Perturbation Quotient: the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude). Shimmer measurements were extracted using Praat.

3.4.2 Phonetic features

Vowel Space Area: Vowel space area is operationalized as the area of the quadrilateral bounded by the centroids of the /i, æ, u, a/ vowel clouds for a given speech segment in Lobanov normalized space (Lobanov, 1971).

Vowel Space Dispersion: Vowel space dispersion is operationalized as the mean pairwise distance between vowel cloud centroids for a given speech segment in Lobanov normalized space.

Vowel Dynamicity: Vowel dynamicity is operationalized as the mean Euclidean distance from the onset to midpoint to off-glide of all vowels within the speech segment in Lobanov normalized space.

Vowel F1 onset: F1 onset is operationalized as the mean value of F1 25% of the way through the vowel duration in Lobanov normalized space of a given vowel type over a speech segment. The feature vector for each speech segment includes a measure for F1 onset for each of the 15 vowels examined and plotted in chapter 4.

Vowel F1 midpoint: F1 midpoint is operationalized as the mean value of F1 50% of the way through the vowel duration in Lobanov normalized space of a given vowel type over a speech segment. The feature vector for each speech segment includes a measure for F1 midpoint for each of the 15 vowels examined and plotted in chapter 4.

Vowel F1 offglide: F1 offglide is operationalized as the mean value of F1 75% of the way through the vowel duration in Lobanov normalized space of a given vowel type over a speech segment. The feature vector for each speech segment includes a measure for F1 offglide for each of the 15 vowels examined and plotted in chapter 4.

Vowel F2 onset: F2 onset is operationalized as the mean value of F2 25% of the way through the vowel duration in Lobanov normalized space of a given vowel type over a speech segment. The feature vector for each speech segment includes a measure for F2 onset for each of the 15 vowels examined and plotted in chapter 4.

Vowel F2 midpoint: F2 midpoint is operationalized as the mean value of F2 50% of the way through the vowel duration in Lobanov normalized space of a given vowel type over a speech segment. The feature vector for each speech segment includes a measure for F2 midpoint for each of the 15 vowels examined and plotted in chapter 4.

Vowel F2 offglide: F2 offglide is operationalized as the mean value of F2 75% of the

way through the vowel duration in Lobanov normalized space of a given vowel type over a speech segment. The feature vector for each speech segment includes a measure for F2 offglide for each of the 15 vowels examined and plotted in chapter 4.

3.4.3 Lexical features

Quotative Frequency: Quotative frequency is operationalized as the normalized frequency (per 1,000 words) of a given quotative within a speech segment. The feature vector for each speech segment includes a measure of normalized frequency for each of the following quotatives: "be all", "be like", "say", and "go." It should be noted that, as the corpus transcripts do not explicitly code reported speech as such, the normalized frequencies used in the feature vector are frequencies of all occurrences of these (lemmatized) tokens. This likely artificially boosts the frequency of the polysemous quotatives, particularly "go." Frequency of the construction "be like" may also be somewhat inflated due to the discourse marker usage of "like."

Modal Frequency: Modal frequency is operationalized as the normalized frequency (per 1,000 words) of a given modal within a speech segment. The feature vector for each speech segment includes a measure of normalized frequency for each of the following modal constructions: "will," "would," "shall," "should," "may," "might," "can," "could," "ought," "must," "going to," "have to," and "need to." These modal constructions are drawn from Barbieri (2008)

Intensifier Frequency: Intensifier frequency is operationalized as the normalized frequency (per 1,000 words) of a given intensifier within a speech segment. Intensifier is defined for the purposes of this dissertation as a lexeme whose lemma matches the form of one of the top 13 most common intensifiers identified by Barbieri (2008) and whose head is an adjective or an adverb. The feature vector for each speech segment includes

a measure of normalized frequency for each of these lexemes. The full list of intensifiers may be found in appendix D.

Discourse Marker Frequency: Discourse marker frequency is operationalized as the normalized frequency (per 1,000 words) of a given discourse marker within a speech segment. As with intensifiers, the set of discourse markers examined in this dissertation is drawn from Barbieri (2008). The feature vector for each speech segment includes a measure of normalized frequency for each member of this set. The full list of discourse markers used may be found in appendix D.

Taboo Frequency: Taboo frequency is operationalized as the cumulative normalized frequency (per 1,000 words) of all lexemes within a speech segment which are designated as "taboo." The set of lexemes considered "taboo" for the purpose of this dissertation is drawn from Lancker and Cummings (1999) and includes profanity, racial slurs, genitalia terms, sexual slurs, and so on. The full list of lexemes considered as taboo for the purposes of this dissertation may be found in appendix D.

Politeness Frequency: Politeness frequency is operationalized as the cumulative normalized frequency (per 1,000 words) of all lexemes within a speech segment which are designated as "polite." The set of lexemes considered "polite" for the purpose of this dissertation is based on the politeness speech act formulae defined in the Longman Grammar of Spoken and Written English (Biber et al., 1999). The full set of politeness terms may be found in appendix D.

Polarity: Segment Polarity is operationalized as a float between -1.0 and 1.0, with higher numbers corresponding to positive sentiment and lower numbers corresponding to negative sentiment. Polarity is implemented via the TextBlob package in Python, which calculates polarity over the segment as a whole via rule-based combinatorial transformations of polarity scores assigned to individual words within the segment.

Polarity scores used here for individual words come from the default polarity score table of the TextBlob package. It should be noted that this is not a particularly sophisticated implementation of segment polarity, and may be interpreted with some caution.[13]

Subjectivity: Segment subjectivity is operationalized as a float between 0.0 and 1.0, 0.0 corresponding to objective statements (i.e. factual information) and 1.0 corresponding to subjective statements (i.e. personal opinion, emotion, judgment, etc.). Subjectivity is implemented via the TextBlob package in Python, which calculates a subjectivity score over the segment as a whole much the same as it calculates polarity: via rule-based combinatorial transformations of subjectivity scores assigned to individual words within the segment. Subjectivity scores used here for individual words come from the default subjectivity score table of the TextBlob package. As an example, the objective statement "the temperature is 95 degrees" is assigned a subjectivity score from TextBlob of 0.0, whereas the subjective statement "it feels hot" is assigned a subjectivity score of 0.85. As with polarity, this is not a particularly sophisticated implementation of subjectivity and should be interpreted with some caution.

Average word Length: Word length is operationalized as the mean number of syllables contained in each token in a given speech segment. Syllable boundaries within a given word are detected using the US English CMU pronouncing dictionary.[14]

Speech Rate: Speech rate is operationalized as the average number of tokens (words) per minute within a given speech segment.

Informative ngrams: For each of the five social traits of focus, the top 2,000 ngrams

[13]Polarity and Subjectivity are included in this feature set for largely exploratory purposes. To my knowledge these features have not been explored in the sociolinguistic literature as features capable of distinguishing between social categories. However, polarity and subjectivity are the subject of much computational interest, and it is not implausible that they might be indicative of sociodemographic distinctions, hence their inclusion here.

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

most informative for distinguishing between trait classes were determined via information gain ranking using a binary presence/absence coding schema. The max rank of ngram considered was 2 (i.e. all unique unigrams and bigrams present within the corpus were considered as candidates). These 2,000 most informative ngrams for each specific trait were then included as features (using the same binary presence/absence coding schema) for each speech segment on tasks aimed at predicting that particular trait (e.g. for single task learning models focused on age prediction, the feature vectors delivered to the models included all previously mentioned features as well as the presence/absence of each of the top 2000 ngrams for age). A binary coding schema was chosen for the ngram features based on preliminary testing that showed degraded performance for models trained on raw frequency and normalized frequency ngram features as compared to binary (presence/absence) ngram features. The top 2,000 ngrams were chosen from the combined set of possible unigrams and bigrams due to preliminary testing which suggested improved performance for models trained on ngram features selected in this manner over models trained on ngram features selected from exclusively either the set of unigrams or the set of bigrams. Inclusion of informative ngrams as features was motivated by Gillick (2010), who demonstrated excellent classification results on similar speaker profiling tasks using similar informative ngram features.

For a detailed examination of the distribution of each of these features with respect to social traits, please refer to chapter 4. All methodological points concerning the design and implementation of the models trained on the data and features discussed in this chapter can be found in chapters 5 and 6.

Chapter 4: Data Exploration

This chapter examines each of the extracted predictor features with respect to the five sociodemographic traits of focus: Sex, Ethnicity, Age, Region, and Education. Examining these predictors prior to and independently of modeling allows for a better sense of which predictors are likely to be important for which sociodemographic traits and informs discussion of the feature importance results presented in chapter 7. Trends observed in visual examinations of the predictive features are confirmed by linear mixed effects models where appropriate. Unless otherwise specified, the linear mixed effects models reported below treat the specific feature in question as the dependent variable, the social trait in question as a fixed effect, and include a random intercept for speaker ID. Refer to Appendix C for a comprehensive list of results from all linear mixed effects models fit to the data throughout the course of this chapter.

As mentioned in chapter 3, the 10,010 conversational speech segments which form the unit of analysis in this dissertation are excerpted from speech data gathered from 669 different speakers. In most cases this chapter will examine feature distribution with respect to sociodemographic traits at the segment level (i.e. 10,010 data points, one for each speech segment) rather than the speaker level (i.e. 669 data points– one for each speaker), as this is the level of granularity at which modeling will occur. All sociodemographic trait overview sections however are presented at both the segment and speaker levels.

4.1 Sex

4.1.1 Overall sex distribution

Prior to examining individual predictors with respect to sex, it's important to understand the overall distribution of sex throughout the data-set as a whole.

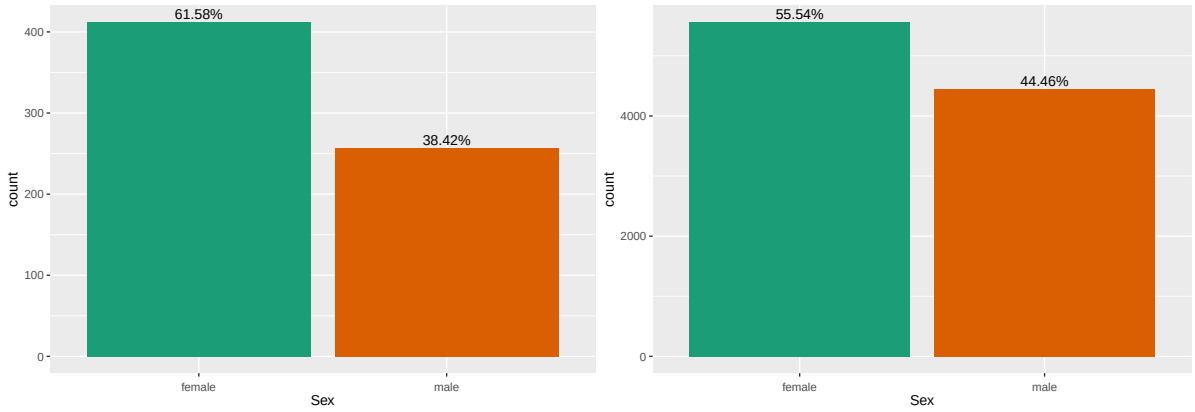

Figure 4.1: Sex breakdown by speakers (left) and segments (right)

Somewhat of a class imbalance exists at the speaker level, with generally more female representation than male representation. As seen in figure 4.1, of the 669 speakers included in the data-set, 38.42% are male and 61.58% are female.

This imbalance is somewhat alleviated when looking at the numbers broken down by segment rather than by speaker. Of the 10,010 conversational speech segments analyzed, 44.46% are male and 55.54% are female. This speaks to the fact that though there is a large imbalance in the total number of unique female and male speakers represented in the data, male speakers on average took part in a greater number of telephone conversations during the collection of the original NIST data-set than did female speakers.

4.1.2 Acoustic variables

Below is an examination of the acoustic predictors extracted from the data with respect to sex.

4.1.2.1 Harmonic to noise ratio (HNR)

Figure 4.2 shows a clear difference in HNR between males and females in the corpus, with females producing higher HNR values than males. A linear mixed effects model fit to the data confirms a moderately strong and statistically significant effect of speaker sex on HNR ($\eta^2 = 0.208$, $p < 0.001$), suggesting that HNR will indeed be a useful feature to include when modeling speaker sex.

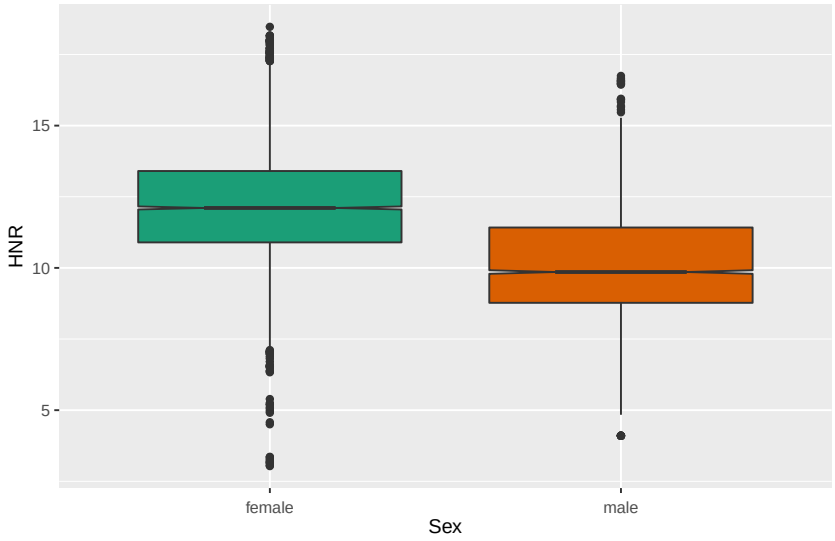

Figure 4.2: Sex breakdown of HNR

4.1.2.2 Jitter

All three measures of jitter (absolute, five-point period perturbation quotient, relative average perturbation) show a clear differentiation between the sexes as seen in figure

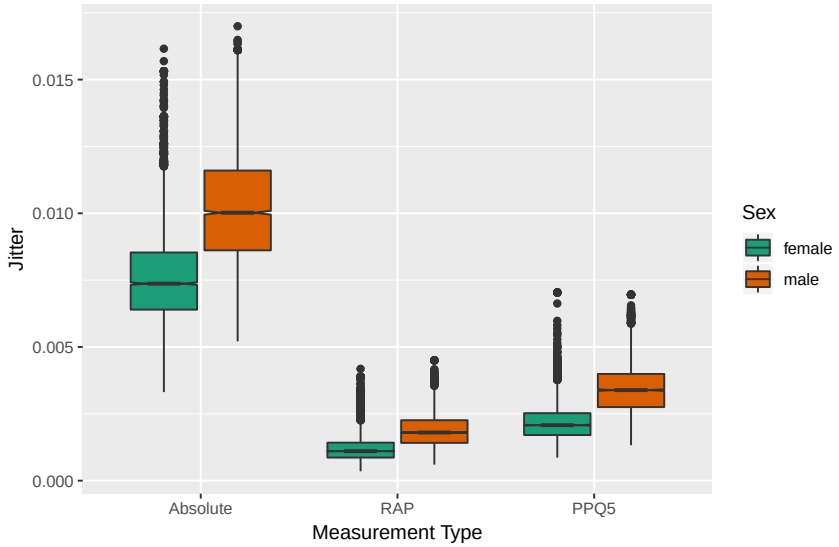

Figure 4.3: Sex differences in jitter measurements at the segment level

4.3, with the five-point period perturbation quotient (ppq5) appearing to show slightly more differentiation than the relative average perturbation (rap) and absolute measures. Linear mixed effects models fit to the data confirm moderate to strong, statistically significant effects of speaker sex on all three measures of jitter (Absolute: $\eta^2 = 0.346$, $p < 0.001$; RAP: $\eta^2 = 0.296$, $p < 0.001$; PPQ5: $\eta^2 = 0.405$, $p < 0.001$). That male speakers would exhibit higher jitter values is in line with expectations, and suggests that jitter would be a meaningful predictor to include when modeling sex. As all three measures of jitter are highly correlated with one another (Pearson's $r = 0.87$), only one of these jitter measures should be used in modeling.

4.1.2.3 Shimmer

As with jitter, both shimmer measures (absolute, three-point amplitude perturbation quotient) appear to show a clear difference between the sexes, with absolute shimmer exhibiting slightly more differentiation than the three-point amplitude perturbation quotient (apq3) measure. Moderately strong and statistically significant effects of speaker

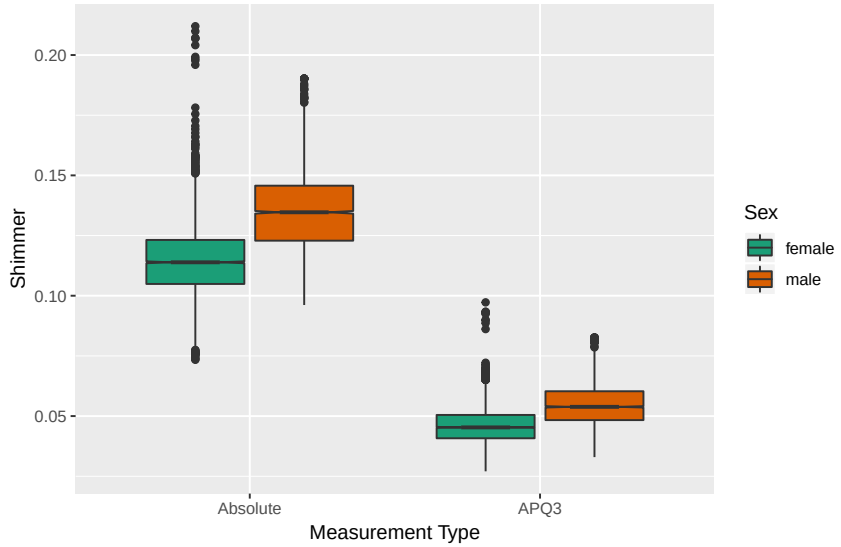

Figure 4.4: Sex differences in shimmer at the segment level

sex on both measurements of shimmer are confirmed by linear mixed effects models fit to the data (Absolute: $\eta^2 = 0.321$, $p < 0.001$; APQ3: $\eta^2 = 0.212$, $p < 0.001$). The direction of the effect is again expected, men exhibiting higher shimmer values than women, and as the two measures are highly correlated ($r = 0.91$) only one should be included as a predictor during the modeling stage.

4.1.2.4 Pitch

Three measures of pitch were observed: the maximum pitch reached throughout a conversation, the mean pitch sustained throughout a conversation, and the minimum pitch reached throughout a conversation. Of the three, mean pitch appears to most strongly differentiate male from female speakers. Linear mixed effects models fit to the data confirm significant differences between the sexes for all three measures, though effect size for mean pitch is particularly strong and far outweighs the effect sizes for maximum and minimum pitch (pitch_min: $\eta^2 = 0.023$, $p < 0.001$; pitch_max: $\eta^2 = 0.016$, $p < 0.001$, pitch_mean: $\eta^2 = 0.574$, $p < 0.001$).

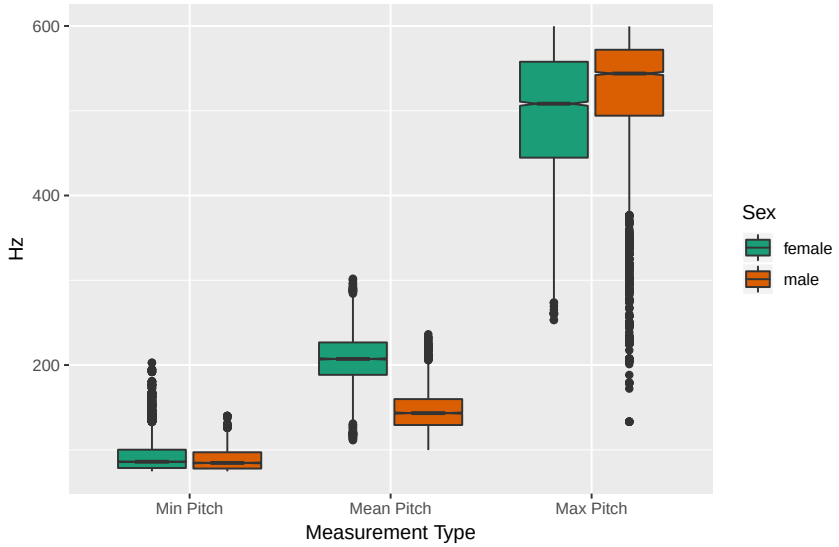

Figure 4.5: Sex differences in pitch at the segment level

Interestingly, while it is to be expected that males on average have lower mean and minimum pitch values than females, figure 4.5 appears to show that males, counter to the literature and intuition, also exhibit a slightly higher maximum pitch in this particular data-set. This may be a measurement error. Maximum and minimum pitch values reached during a conversation simply represent the pitch at one particular point, and thus are more susceptible to momentary inaccuracies by the pitch detection engine of Praat. Mean pitch, being an average over the entire conversation, is less susceptible to momentary erroneous pitch readings. Because of the nature of the max and min pitch collection procedure, these values should be viewed with some suspicion.

4.1.3 Phonetic variables

4.1.3.1 Vowel space

Three features relating to vowel space were extracted for this analysis: vowel space area, vowel space dispersion, and vowel dynamicity.

As figure 4.6 shows, males appear to have a larger vowel space on average than do females. A mixed effects model fit to the data confirms a moderate, statistically significant effect of speaker sex with respect to this variable ($\eta^2 = 0.161$, $p < 0.001$).

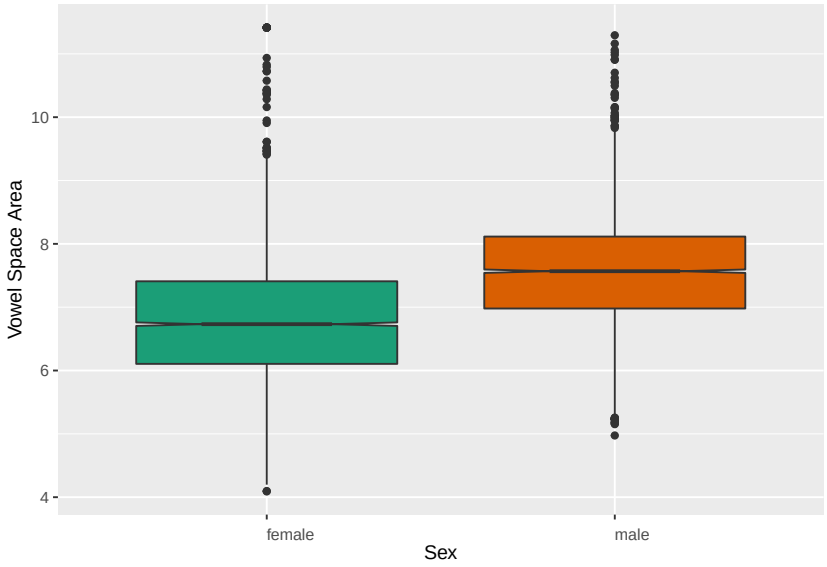

Figure 4.6: Sex differences in vowel space area

Likewise, and somewhat relatedly, figure 4.7 suggests that males also tend to exhibit wider vowel dispersion on average in this data-set than do females. Again, a mixed effects model fit to the data confirms a moderate, statistically significant effect of speaker sex with respect to this variable ($\eta^2 = 0.195$, $p < 0.001$).

The sex differences exhibited in figures 4.6 and 4.7 are interesting, as most literature on this topic suggests the opposite should hold true– namely, females are hypothesized on average to exhibit larger vowel space area and wider vowel dispersion than males during conversational speech, as discussed in chapter 2. This is reminiscent of the counter-intuitive finding discussed in section 4.1.2.4 that males in this data-set tend also to exhibit a higher maximum pitch and wider pitch range than do females. Recall that this data-set is comprised entirely of telephone conversations, whereas most literature on sex differences in acoustic speech characteristics in the literature are based on in-person, conversational speech data or speech data recorded in a laboratory setting.

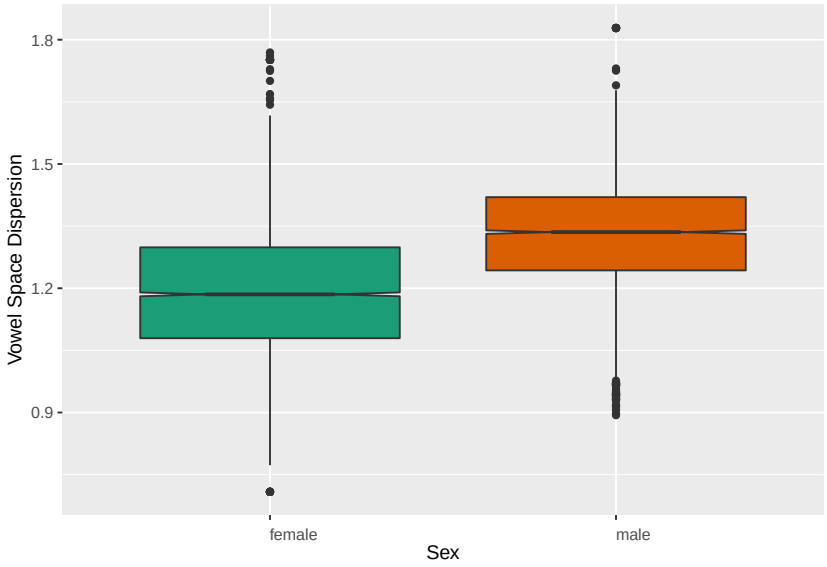

Figure 4.7: Sex differences in vowel space dispersion

It's possible that differences between measures observed in this data-set and expectations based on the literature stem from this difference in genre and/or recording medium. Some work has been done investigating the effects of the telephone medium on acoustic/phonetic properties of speech (e.g. Künzel, 2001), but I'm unaware of any such findings that would account for the particular patterns observed here.

Unlike vowel space area and dispersion, figure 4.8 appears to show no clear difference between sexes when it comes to vowel dynamicity. This is largely expected, as vowel dynamicity is not hypothesized in the literature to exhibit a sex difference.

4.1.3.2 Vowel positions

The vowel space plots in figure 4.9 and throughout the rest of this chapter show Lobanov normalized values for F1 and F2 at three measurement points for each vowel: 25% of the way through the vowel trajectory (onset), 50% of the way through the vowel trajectory (midpoint), and 75% of the way through the vowel trajectory (offglide). As mentioned in chapter 2, measuring vowels at multiple points throughout the trajectory

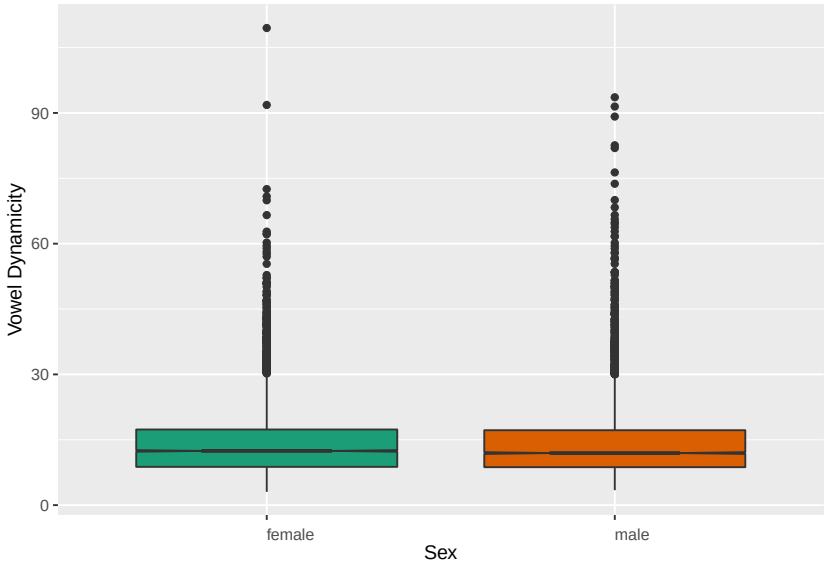

Figure 4.8: Sex differences in vowel dynamicity

rather than taking a single (midpoint) measurement for each vowel increases accuracy in class distinctions for some sociodemographic categories, and is necessary in order to obtain a measure of vowel dynamicity. Onsets are labeled with a square symbol, midpoints with a circle, and offglides with a triangular arrow pointing in the direction of the continuation. The measurements shown in these vowel plots represent the average measurement at that point for that vowel for that particular grouping variable. E.g. the F1 and F2 values for the onset of the UW diphthong for males in figure 4.9 are calculated by first averaging the F1 and F2 measurements 25% of the way through each UW vowel for each segment, and then taking the mean F1 and F2 for the onset of UW across all male segments.

Examining the plot of the vowel trajectories for males and females in figure 4.9 reflects the findings above of generally smaller vowel space area and narrower vowel dispersion for females than for males. This difference appears to primarily be motivated by a narrower range along F2 for females as compared to males. The greatest difference between the sexes appears to be in realizations of front vowels, which are produced much further back in the vocal tract than those of males. In contrast to F2, females

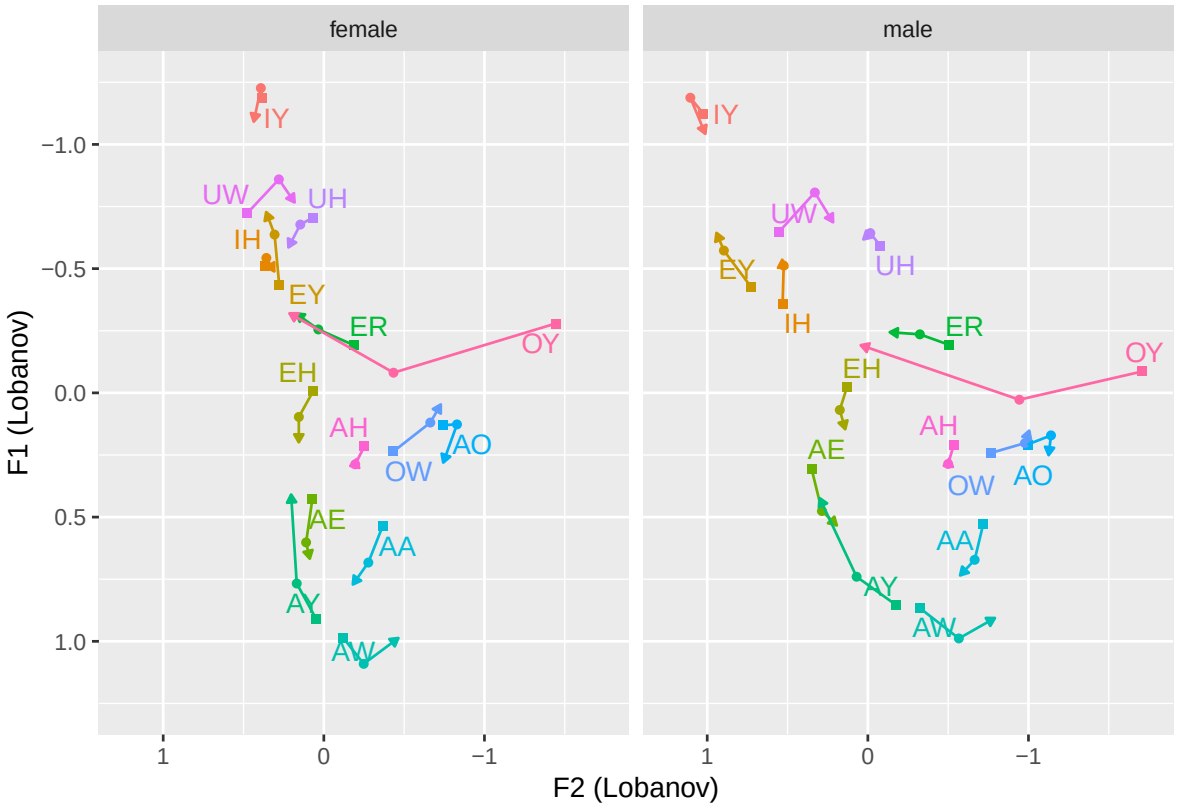

Figure 4.9: Vowel trajectory differences between sexes

appear to exhibit a slightly wider range along F1 as compared to males– the high vowels being realized slightly higher, and the low vowels being realized slightly lower. Refer to appendix C for a comprehensive list of effect size and significance of speaker sex on realizations for each individual vowel.

4.1.4 Lexical features

4.1.4.1 Quotatives

Kword frequency (i.e. frequency normalized per 1,000 words) was extracted for four different quotatives: "be all", "be like", "say", and "go," reported in figure 4.10. It should be noted that, as the data-set does not explicitly code reported speech as such, the kword frequencies reported here are frequencies of all occurrences of these constructions rather than just the instances in which the construction in question was used in a definitively quotative context. Figure 4.10 shows the average kword frequency for males and females for each quotative observed.

Females appear to use quotatives at a slightly higher frequency overall in this data set than do males. Though the difference in production rates of "be all" and "go" appear negligible, there does appear to be some sex-based differentiation for production rates of "say" and "be like." Linear mixed effects models fit to the data however do not find these differences to be statistically significant, nor does a linear mixed effects model fit to the data find a significant difference between the sexes in the total frequency of quotatives (i.e. quotative rate when instances of all four quotatives are summed). This is somewhat odd in light of the research discussed in chapter 2 which reports significant sex-based variation in rates of quotative production. The lack of a meaningful sex effect here may be a result of the inability to distinguish instances of these constructions in

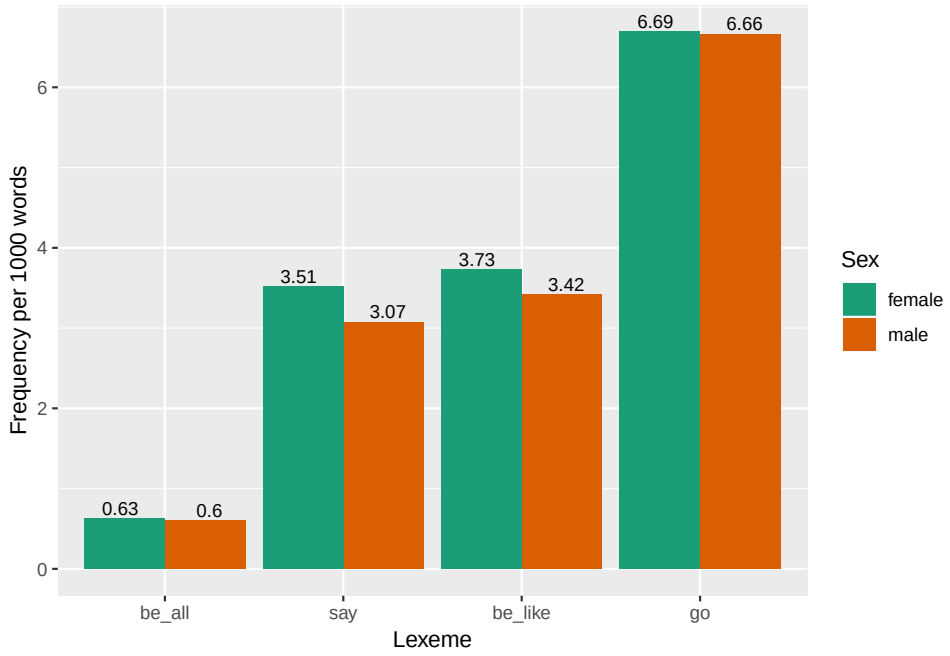

Figure 4.10: Quotative usage by sex

quotative contexts from non-quotative contexts in this particular corpus.

4.1.4.2 Modals

As with quotatives, it appears from figure 4.11 that females tend to produce modals at a slightly higher rate overall than do males. This pattern holds true when examining production rates for most individual modal verbs as well, with the exception of "should" and "could." Linear mixed effects models including a random intercept for speaker ID however do not find significant differences between the sexes in terms of frequency of production of any of the modal constructions examined. As with quotatives, this runs counter to the literature discussed in chapter 2, which reports significant sex-based differences in frequency rates of modal constructions.

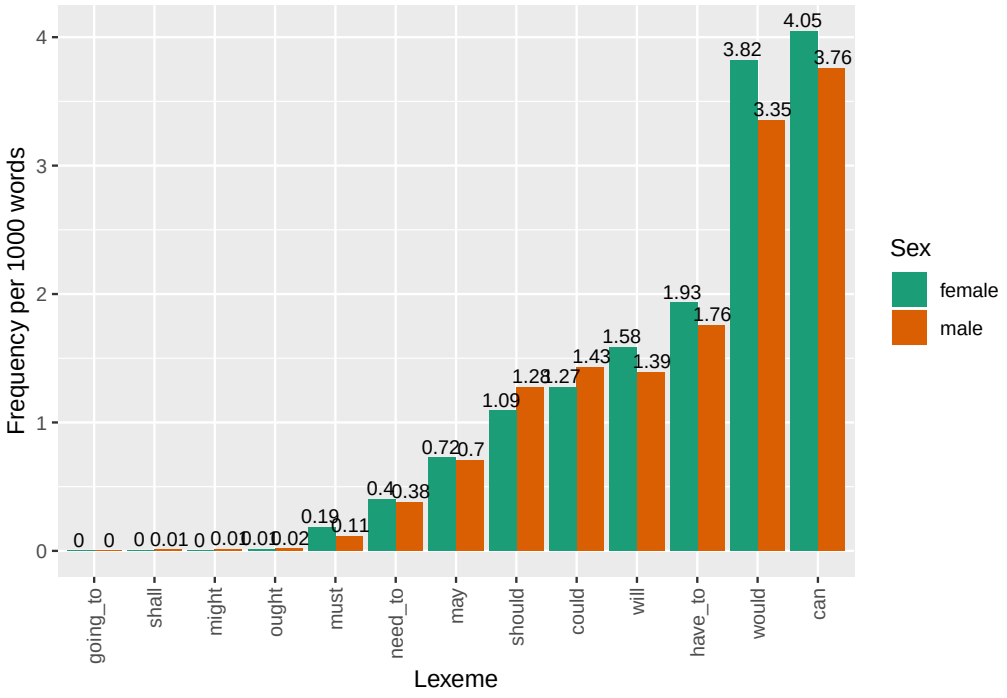

Figure 4.11: Modal usage by sex

4.1.4.3 Intensifiers

It appears from figure 4.12 that females tend on average to produce intensifiers at a slightly higher rate than do males. This holds true for most individual intensifiers examined as well, with the notable exception of "pretty," which appears to be preferred by males. Linear mixed effects models fit to the data indicate weak yet significant effects of speaker sex on frequency rates of "very" ($\eta^2 = 0.006$, $p < 0.05$), "so" ($\eta^2 = 0.007$, $p < 0.05$) and "pretty" ($\eta^2 = 0.019$, $p < 0.001$), but none of the other intensifiers examined here.

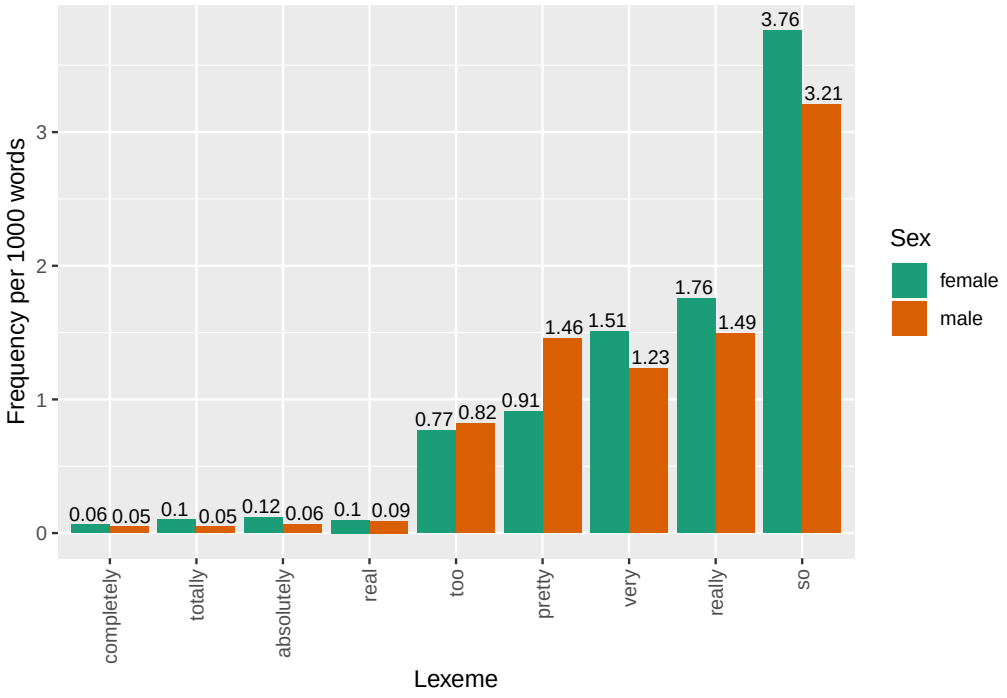

Figure 4.12: Intensifier usage by sex

4.1.4.4 Discourse markers

Figure 4.13 suggests that discourse marker frequency shows little variation between male and female speakers, with the exception of "yeah," which appears to be favored by male speakers. Linear mixed effects models fit to the data find weak yet significant effects of speaker sex for only two of the discourse markers examined: "yeah" ($\eta^2 = 0.014$, $p < 0.01$) and "I know" ($\eta^2 = 0.026$, $p < 0.001$).

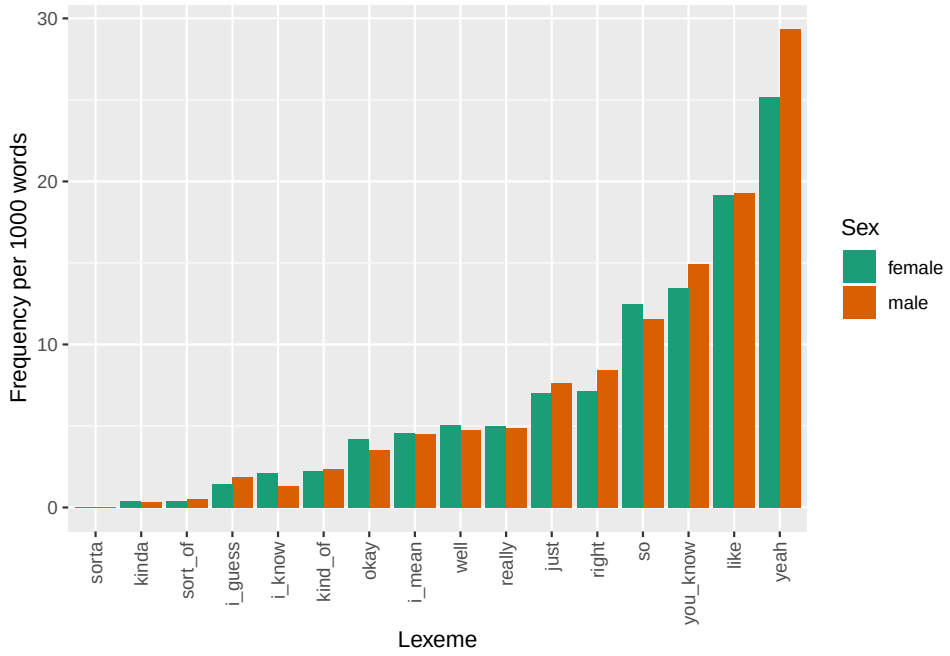

Figure 4.13: Discourse marker usage by sex

4.1.4.5 Pronoun proportions

Figure 4.14 shows the proportion of first, second, and third person pronouns used by male and female speakers in this data-set. There appears to be very little difference in pronoun proportions between the sexes, which is somewhat surprising given the findings described in chapter 2 suggesting that males tend to use first person pronouns at a higher proportion than do females. Linear mixed effects models fit to the data find no significant effect of speaker sex in terms of pronoun proportional usage. A significant (though very weak) effect of speaker sex was however found when comparing the use of the first person singular nominative pronoun "I" individually ($\eta^2 = 0.009$, $p < 0.05$).

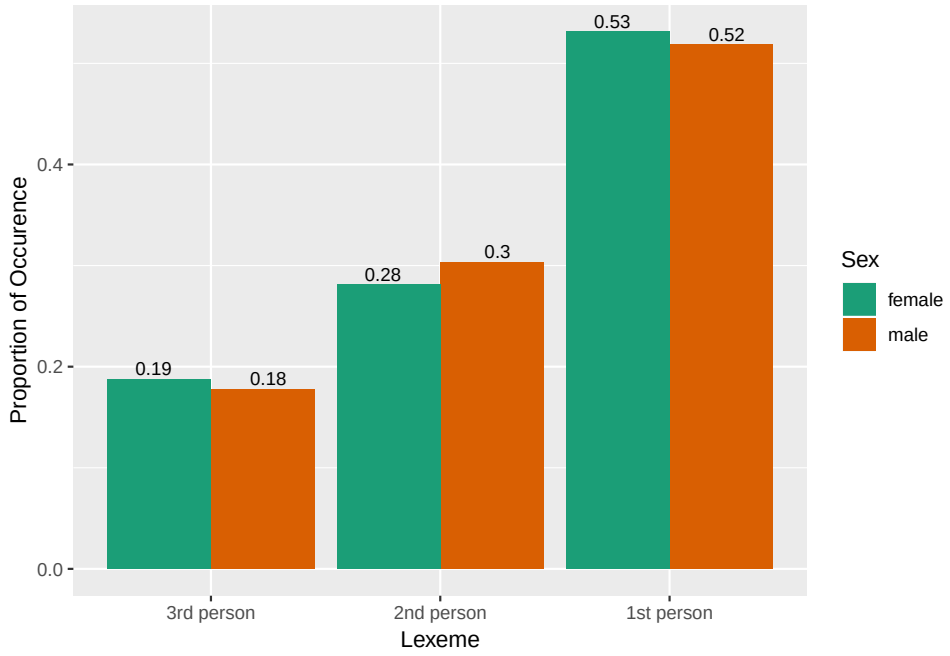

Figure 4.14: Pronoun usage proportions

4.1.4.6 Politeness and taboo frequency

Figure 4.15 shows the average kword frequency of "polite" words and "taboo" words for males and females in this corpus. Words belonging to the "polite" group are selected based on the politeness speech act formulae defined in Biber et al. (1999). Words belonging to the "taboo" group are drawn from Lancker and Cummings (1999). The full set of lexical items included in each of these categories can be found in appendix D.

Though some research suggests that males tend to use taboo words more frequently, no significant difference was found in this corpus between males and females with respect to frequency of taboo language. This may be related to the fact that all speech segments used here come from conversations between strangers. One might expect to see more of a pronounced effect of sex on taboo word frequency in a corpus of more familiar speech wherein use of taboo language is less constrained in general. There

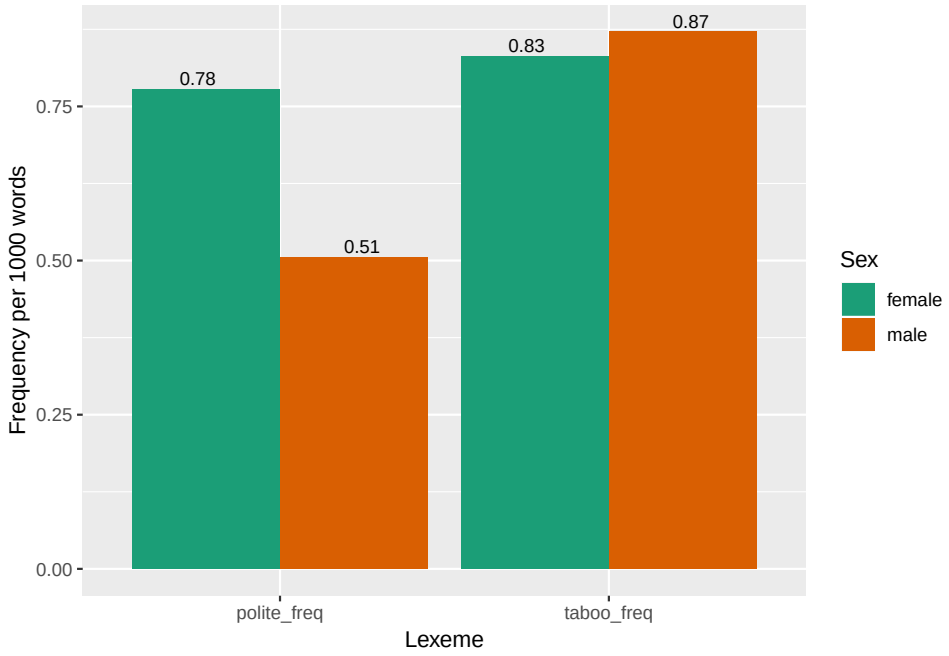

Figure 4.15: Usage of taboo and politeness terms by sex

does appear to be somewhat of a sex difference in frequency of politeness terminology, however this effect appears to be driven largely by one or two individual speakers. A linear mixed effects model fit to the data including a random intercept for speaker ID found the effect of sex on politeness frequency to be quite weak and non-significant ($\eta^2 = 0.005$, $p = 0.07$).

4.1.4.7 Sentiment

Figure 4.16 presents the utterance subjectivity and polarity of speech segments in the corpus grouped by sex. There appears to be little if any difference in either of these metrics based on sex. This is to be expected, as I'm not aware of research suggesting a difference in these metrics between speakers of different sexes.

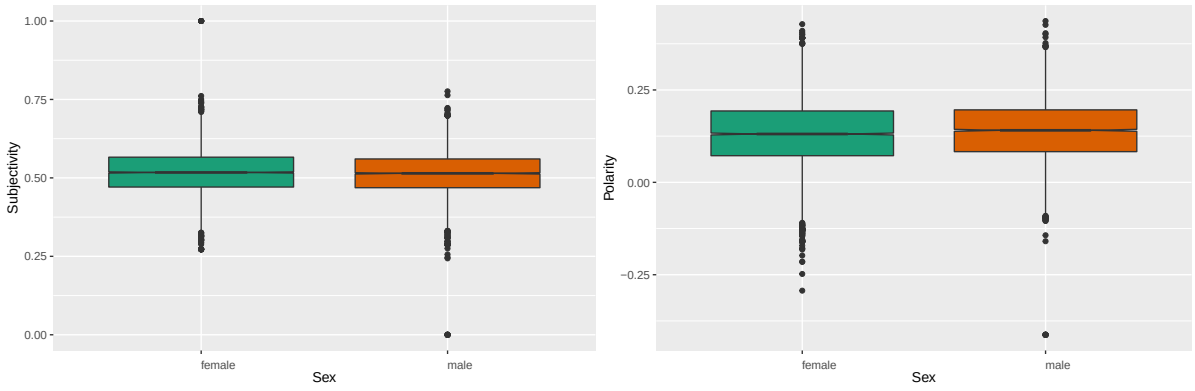

Figure 4.16: Segment subjectivity (left) and polarity (right) by sex

4.1.4.8 Speech rate and word length

Figure 4.17 presents a comparison of the speech rate (tokens per minute) and average word length (in syllables) between male and female speakers in this corpus. There does not appear to be any meaningful difference between the sexes with respect to these predictors– a fact confirmed by linear mixed effects models fit to the data.

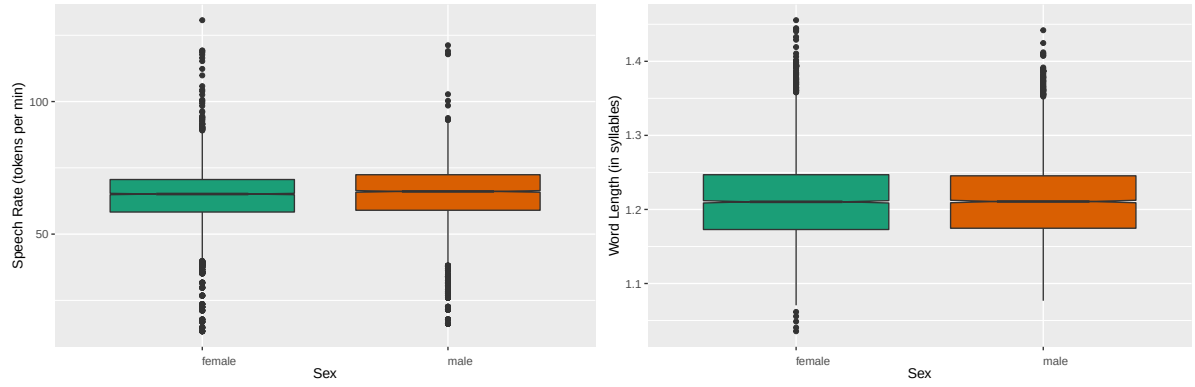

Figure 4.17: Speech rate (left) and word length (right) by sex

4.1.4.9 Top informative ngrams

As detailed in chapter 3, the top 2000 ngrams for each sociodemographic trait of focus– selected from the combined set of all unigrams and bigrams present in the corpus via

Table 4.1: Top 20 informative ngrams for sex

attributes	infogain
my_husband	0.0326438
uh	0.0286200
my_daughter	0.0277263
husband	0.0260696
uh_i	0.0258753
uh_it	0.0237811
ah	0.0219320
uh_you	0.0209853
be_uh	0.0209301
uhhuh_uhhuh	0.0206658
uh_uh	0.0174944
i_um	0.0171691
like_uh	0.0170222
daughter	0.0168330
like_ah	0.0165650
oh_god	0.0156171
you_get	0.0155488
stuff_like	0.0148903
my_wife	0.0143022
know_uh	0.0137681

information gain ranking– were extracted to serve as predictor features for the models presented in chapters 5 and 6. Table 4.1 presents the top 20 ngrams according to information gain for speaker sex. The full 2000 are not shown here for reasons of space.

As the table shows, many of the informative ngrams for sex tend to include interjections such as “uh”, “um”, etc., which in this dataset are heavily favored by male speakers, as well as relation terms (e.g. “wife”, “husband”, “daughter”), which overwhelmingly favor one sex or the other (except for “wife,” ngrams including relation terms tend to favor females).

4.2 Ethnicity

4.2.1 General breakdown

As with sex, it is instructive to examine the overall distribution of ethnicity within the corpus prior to focusing on ethnicity with respect to any one predictor. Figure 4.18 shows the distribution of both speakers and segments for the four ethnic groups considered in this dissertation.

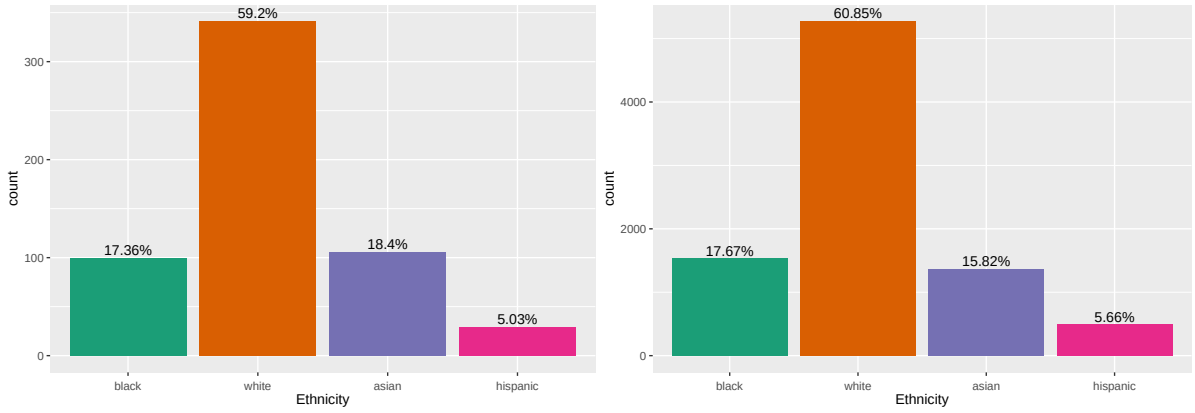

Figure 4.18: breakdown of ethnicity by speakers (left) and segments (right)

There appears to be a large class imbalance in the corpus among these four ethnicities, with very few Hispanic speakers and more than 50% of speakers reporting as White. This class imbalance could be problematic down the line, and may lead to over-prediction of the "White" category unless some steps are taken to address this. Strategies to rectify this class imbalance are discussed in more detail in chapter 5.

The proportion of segments from speakers in each ethnic category is nearly identical to the ethnic proportion among the total number of speakers, indicating that speakers from the four ethnic groups considered here took part in telephone conversations during data collection at roughly similar rates.

4.2.2 Acoustic features

4.2.2.1 Harmonic to noise ratio (HNR)

Though the differences among ethnicities are somewhat slight, figure 4.19 suggests that Asian speakers may have somewhat higher values for HNR than the other three groups on average, and that Hispanic speakers may have somewhat lower values of HNR on average compared with the other three groups.

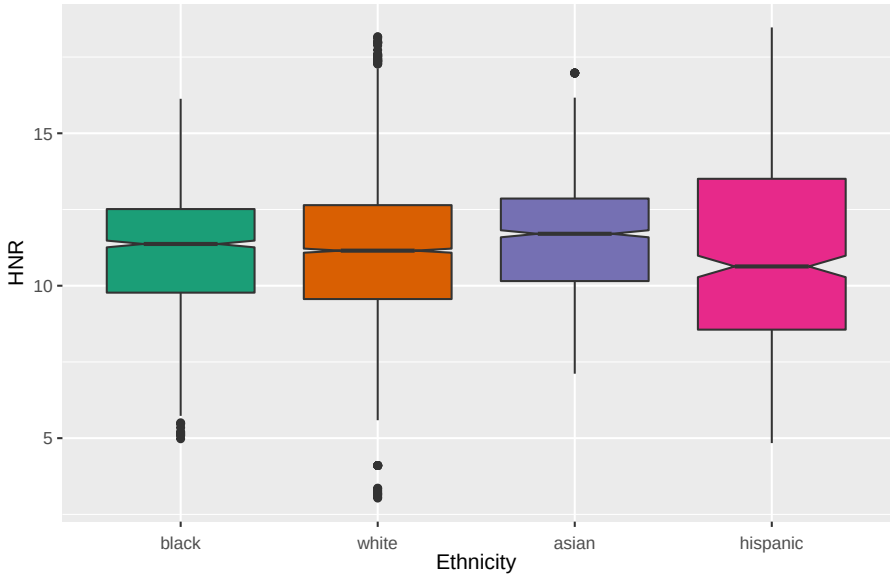

Figure 4.19: Difference in HNR between ethnicities

That Asian speakers would exhibit higher values for HNR than the other four ethnicities is in line with the findings from Newman and Wu (2011), as discussed in chapter 2, that Asian-American speakers used a significantly 'breathier' voice quality than Latino-, African-, and European-American speakers. That Hispanic speakers would exhibit lower values of HNR on average than the other three groups however is not to my knowledge something that has been widely discussed in the literature, although certain styles of Chicano English have been linked to higher-than-average incidences of "creaky voice"" (Mendoza-Denton, 2011), and prototypical creaky voice

generally exhibits lower values of HNR than other types of phonation (Keating et al., 2015). Despite this, a linear mixed effects model found the effect of ethnicity on HNR values to be quite weak and non-significant, suggesting that the variation seen in figure 4.19 may be a result of small sample size for the Asian and Hispanic groups.

4.2.2.2 Jitter

There appears to be very little difference among Asian, White, and Black speakers with respect to jitter. Hispanic speakers however seem to exhibit slightly higher values of jitter for all three measurements. As with HNR, there is no claim in the literature that I know of which would lead us to expect such a difference, and so this may simply be a result of the small sample size for Hispanic speakers within the data-set. In support of this, a linear mixed effects model found effects of ethnicity on jitter values to be extremely weak and non-significant.

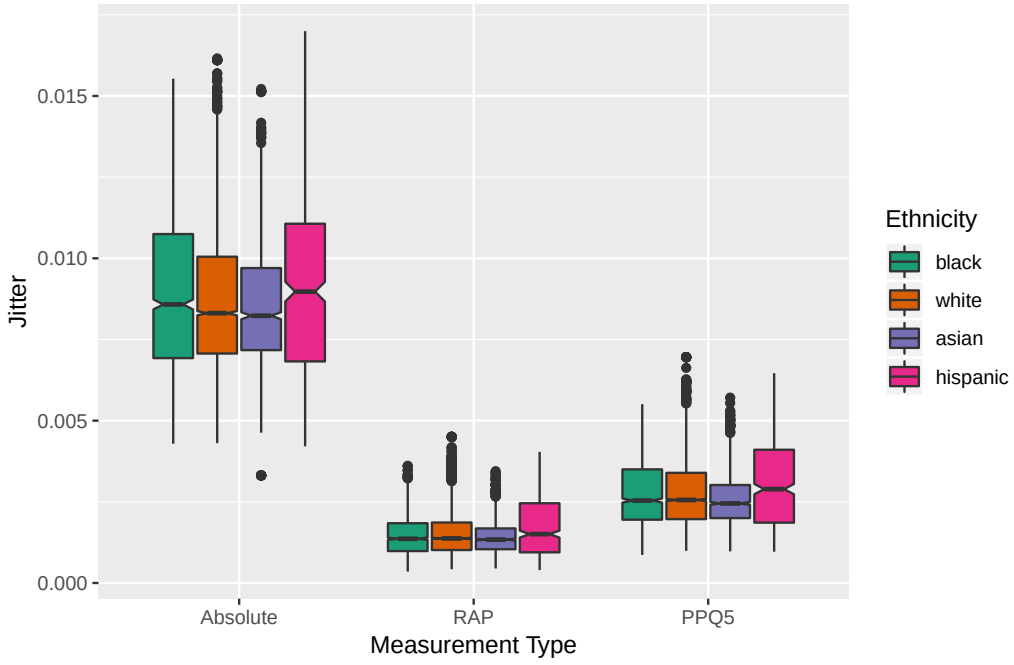

Figure 4.20: Ethnicity differences in jitter measurements at the segment level

4.2.2.3 Shimmer

Figure 4.21 does not appear to show much if any meaningful difference between ethnicities with respect to either absolute or apq3 measures of shimmer. Confirming this, a linear mixed effects model found effects of ethnicity on shimmer values to be extremely weak and non-significant.

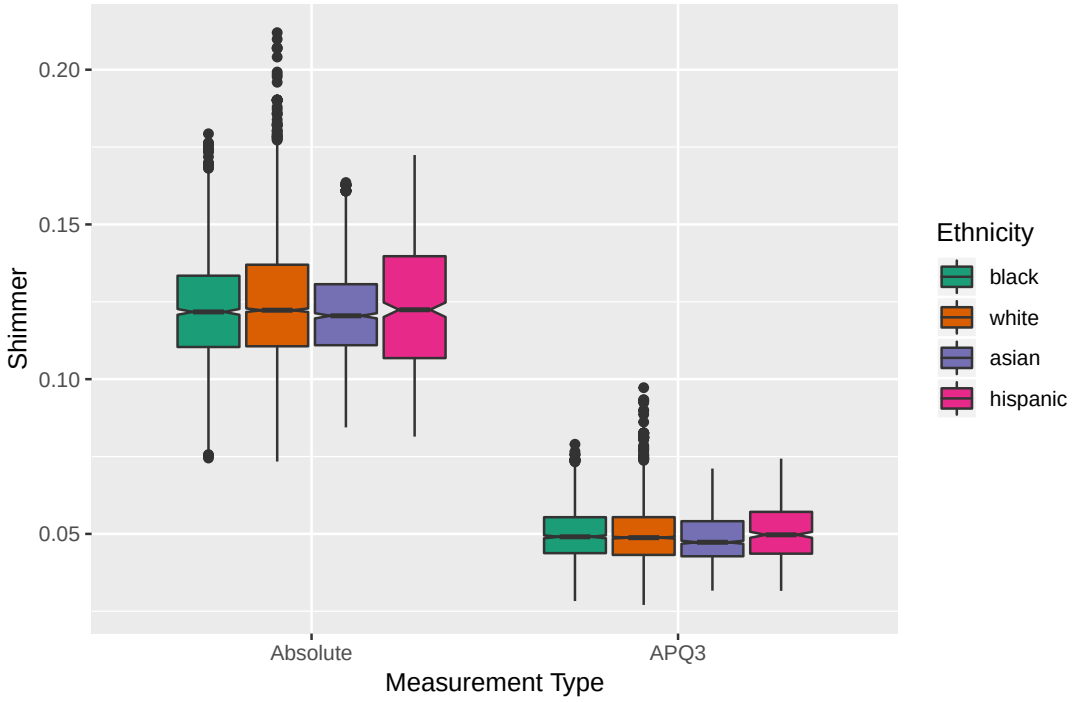

Figure 4.21: Ethnicity differences in shimmer at the segment level

4.2.2.4 Pitch

It appears from figure 4.22 that Hispanic speakers exhibit the highest maximum pitch values and the lowest minimum pitch values, indicating that they may have the widest pitch range of the three ethnic groups. As mentioned above however, maximum and minimum pitch values should be suspect as they are likely prone to measurement errors. Mean pitch values appear to be slightly higher on average for White speakers than for

speakers of the other three ethnic groups. A linear mixed effects model found effects of ethnicity on all three measures of pitch to be weak and non-significant.

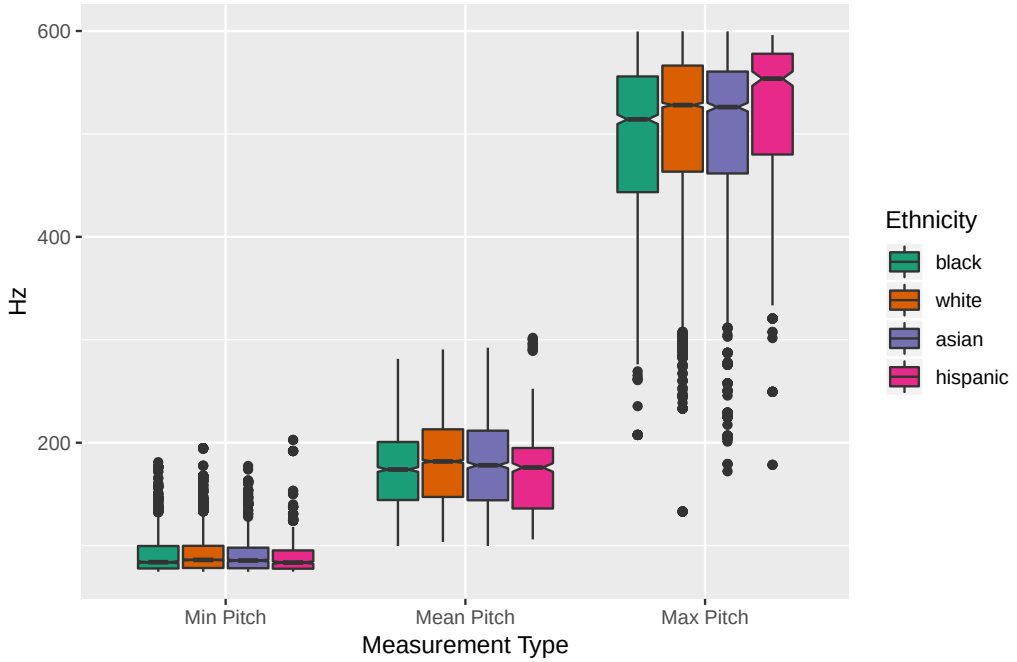

Figure 4.22: Ethnicity differences in pitch at the segment level

4.2.3 Phonetic variables

4.2.3.1 Vowel space

The differences between ethnicities with respect to vowel space in figure 4.23 appear minimal, with Black and Hispanic speakers perhaps exhibiting slightly larger vowel space areas than Asian and White speakers on average. However, a linear mixed effects model fit to the data finds a weak though statistically significant effect of ethnicity on vowel space area ($\eta^2 = 0.018$, $p < 0.05$).

Similarly, figure 4.24 suggests that Black speakers in this data-set tend to have more dispersed vowels on average than the other three groups. A linear mixed effects

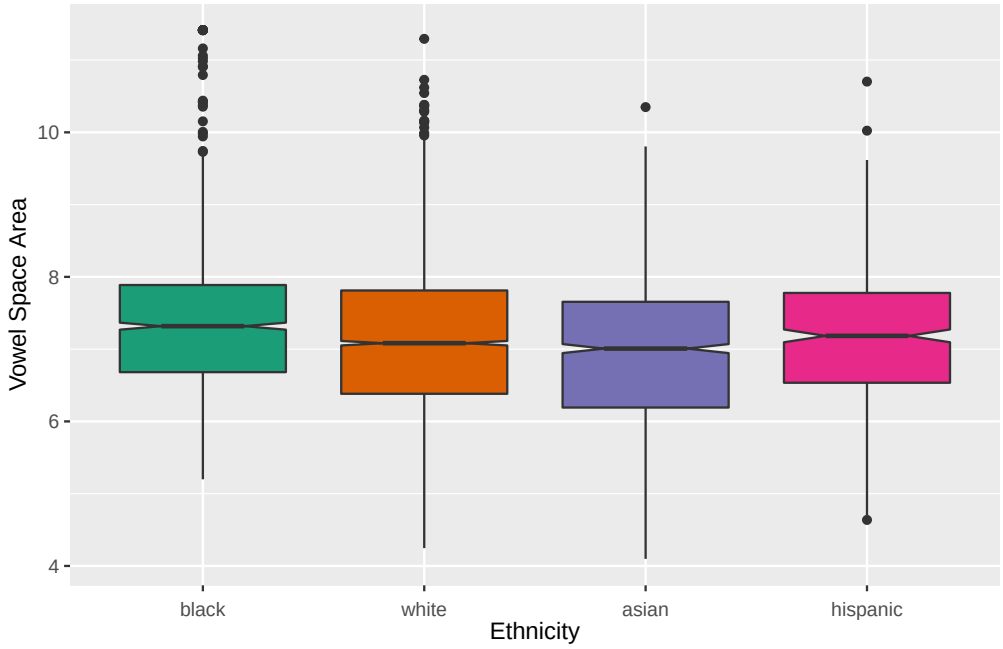

Figure 4.23: Ethnicity differences in vowel space area

model fit to the data find a weak though statistically significant effect of ethnicity on vowel space dispersion ($\eta^2 = 0.022$, $p < 0.01$).

Figure 4.25 shows negligible difference among the ethnic groups in terms of vowel dynamicity, confirmed by a linear mixed effects model.

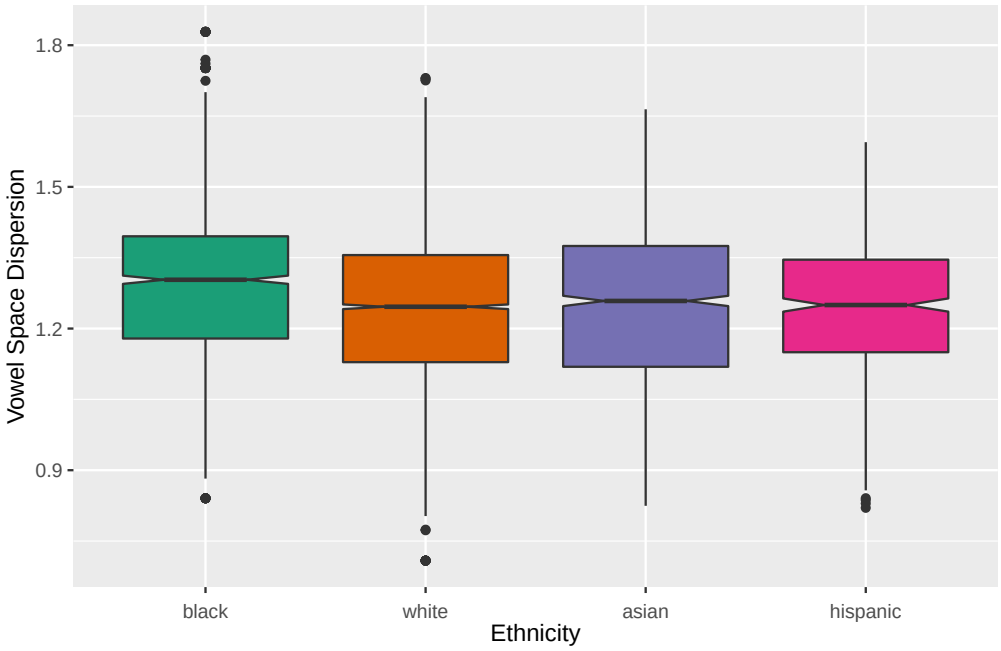

Figure 4.24: Ethnicity differences in vowel space dispersion

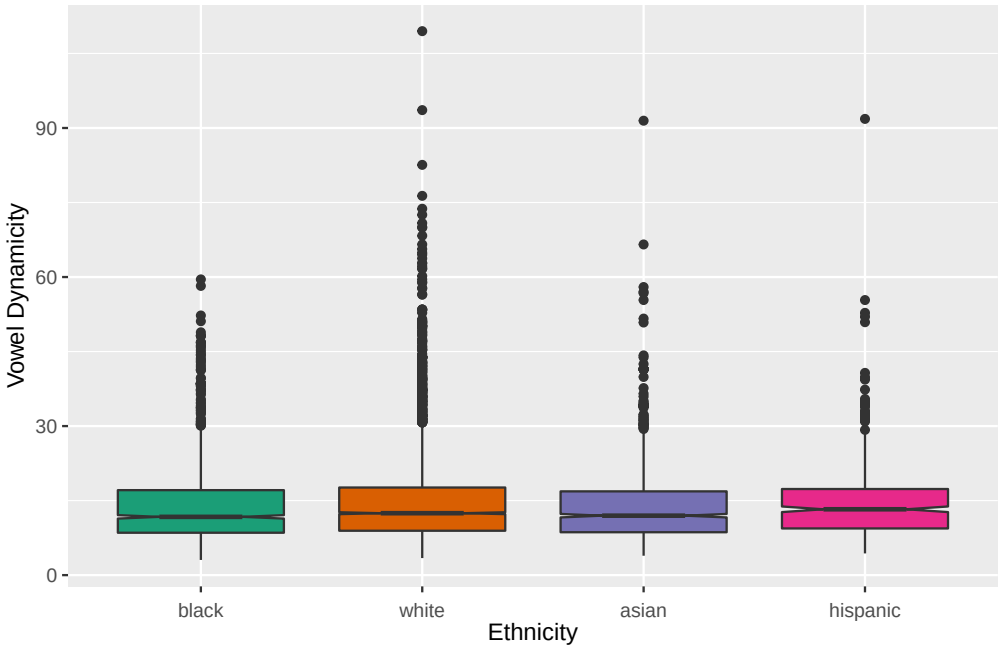

Figure 4.25: Ethnicity differences in vowel dynamicity

4.2.3.2 Vowel positions

Figure 4.26 plots the vowel trajectories for all four ethnic groups in accordance with the plotting conventions laid out above in section 4.1.3.2.

A list of what appear in figure 4.26 to be the salient differences among ethnicities for the vowels analyzed is provided below. Refer to appendix C for a list of those vowel/point combinations on which ethnicity was shown to have a significant effect.

- **IY**: The high front glide IY appears monophthongal and nearly identical among Black, White, and Asian speakers, but Hispanic speakers in this data-set appear to produce IY with a short, downward off-glide.
- **EH**: The mid front vowel EH appears largely similar among Asian, Hispanic, and White speakers. Compared with these three groups, Black speakers appear to exhibit a relatively fronted and somewhat raised realization of EH.
- **AE**: Similarly, Black speakers in this data-set appear to exhibit a relatively fronted and raised realization of AE.
- **AY**: The overall shape and length of the AY glide trajectory appears similar between all groups, but for Black speakers the entire glide trajectory is realized somewhat lower and slightly further back than the other three groups.
- **AW**: The AW diphthong appears similar for Black and White speakers. Hispanic and Asian speakers in comparison to Black and White speakers appear to have a somewhat more backed onset for AW, and Hispanic speakers have a much more backed off glide than either of the other three groups. Asian speakers appear to have a shorter trajectory than the three other groups, and a relatively raised off-glide.
- **UH**: UH for Black and Hispanic speakers appears somewhat fronted and with a longer trajectory than compared to Asian and White speakers.
- **ER**: ER for White and Black speakers appears somewhat backed compared to Asian and Hispanic speakers. Black speakers also appear to exhibit a lower onset and higher off-glide than the other three groups, leading to a more upward tilted trajectory overall.
- **AA**: Black and Hispanic speakers appear to have lower onsets for AA than do Asian and White speakers. Hispanic speakers appear to have a more monophthongal realization of AA than do the other three groups.
- **AO**: AO for White speakers appears slightly more diphthongal than for the other two groups. White and Black speakers exhibit a more backed off-glide for AO than the other two groups.
- **OW**: White speakers appear to have a more fronted onset and longer trajectory for OW than the other three groups. Asian speakers appear to have the most monophthongal version of this glide, with a more fronted off-glide.

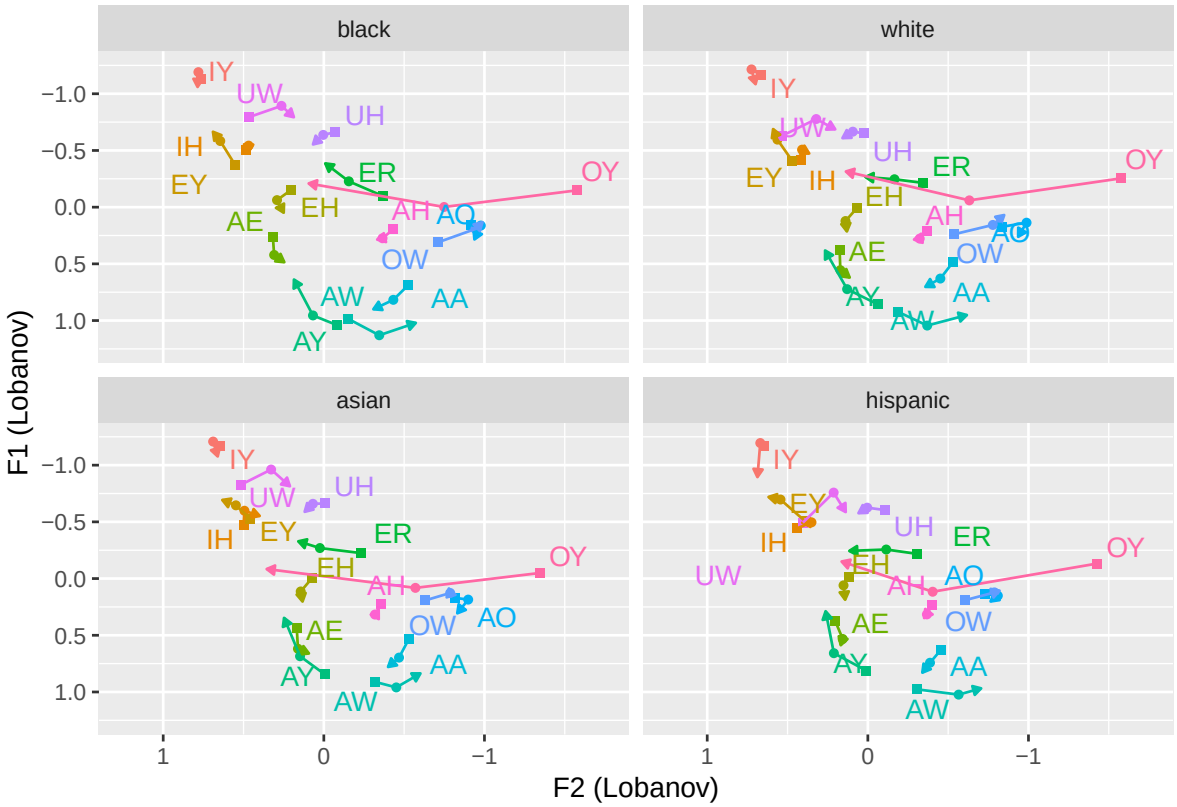

Figure 4.26: Vowel positions and trajectories by ethnicity

4.2.4 Lexical features

4.2.4.1 Quotatives

Figure 4.27 displays the average kword frequency for each of the four quotatives considered in this dissertation for each of the four ethnicities.

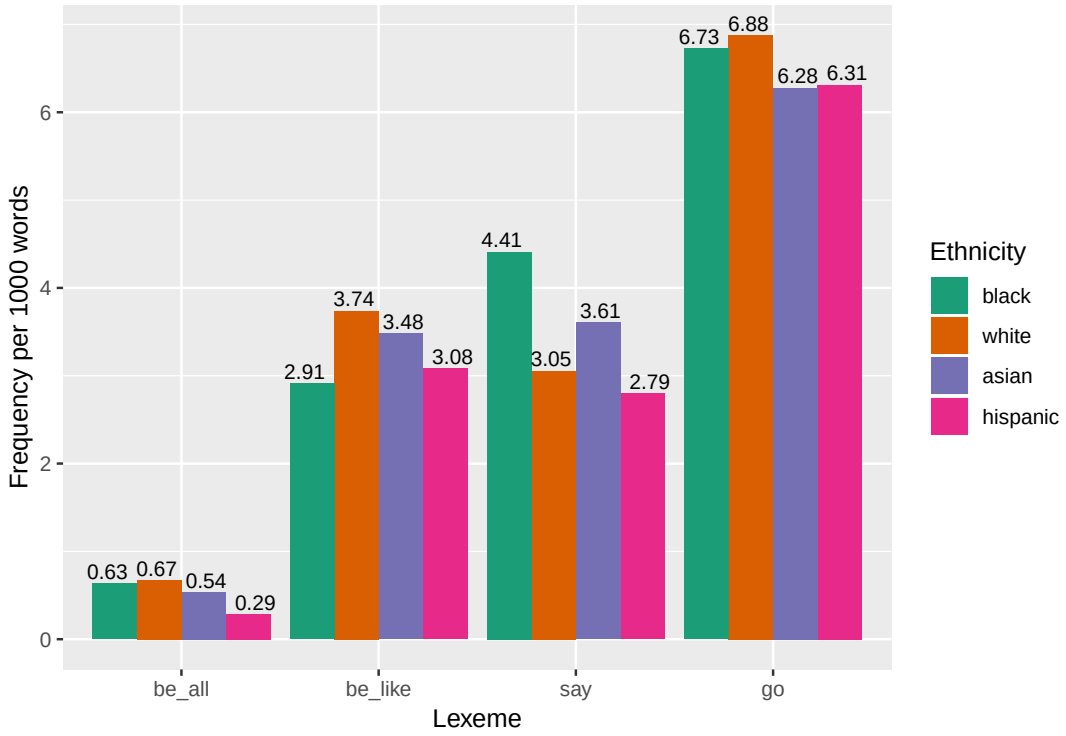

Figure 4.27: Quotative usage by ethnicity

For all four quotatives, Hispanic speakers appear to be in the bottom two in terms of frequency of usage. Black speakers appear to be by far the most frequent users of "say," and the least frequent users of the "be like" quotative. White speakers exhibit the highest usage of "be like" and "go," while exhibiting the relatively low usage of "say." Asian speakers appear to be around the middle of the pack in usage of "be all" and towards the higher end in usage of "be like" and "say." Of the differences suggested by figure 4.27 and discussed here, only the difference between ethnicities with respect to frequency of the "say" quotative was found to be significant ($\eta^2 = 0.023$, $p < 0.01$).

4.2.4.2 Modals

Figure 4.28 shows the kword frequency for the 13 modal constructions considered here. Interestingly, "would", "should," and "have to" all seem to show roughly the same differentiation pattern among the ethnicities, with Asian speakers using these least frequently, Black and Hispanic speakers using these most frequently, and White speakers somewhere in the middle. "Will" shows a somewhat different pattern, used by Black, White, Asian, and Hispanic speakers from most to least frequently, respectively. The other modals do not appear to show much variation.

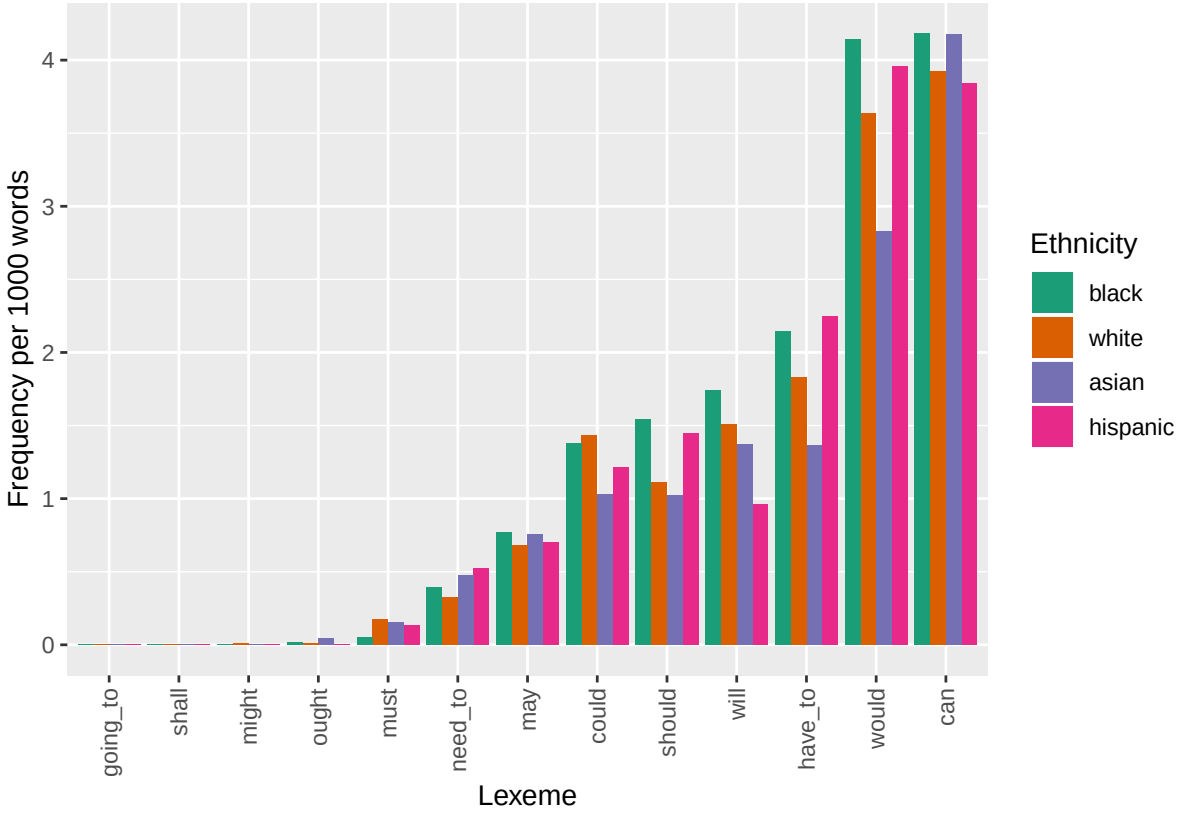

Figure 4.28: Modal usage by ethnicity

Of the differences observed, only the difference with respect to frequency of "would" ($\eta^2 = 0.018$, $p < 0.05$), "could" ($\eta^2 = 0.015$, $p < 0.05$), and "ought" ($\eta^2 = 0.015$, $p < 0.05$) reached significance in linear mixed effects models fit to the data.

4.2.4.3 Intensifiers

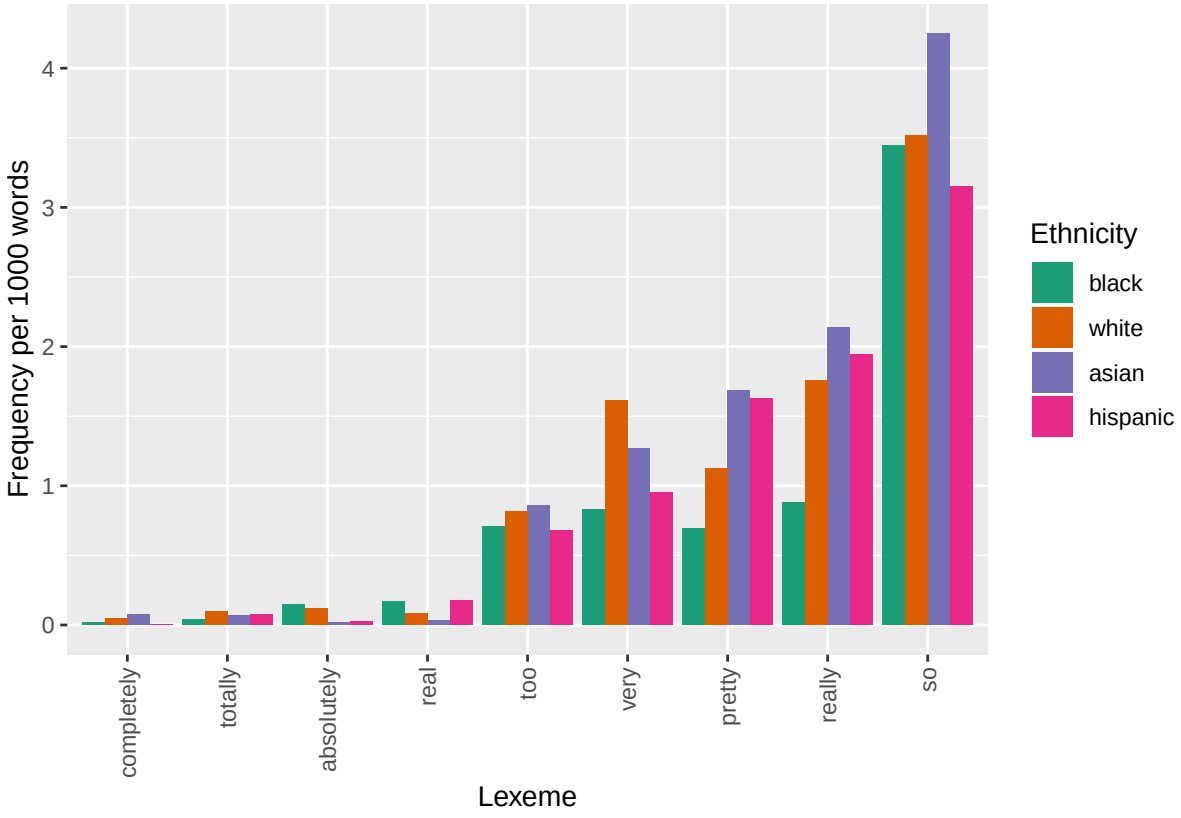

Figure 4.29: Intensifier usage by ethnicity

Among those intensifiers used with any real frequency in the corpus, there appears to be substantial variation in terms of usage among the four ethnic groups. "So" appears to be much more frequent for Asian speakers than the other three groups, whereas "very" is used more frequently by White speakers than speakers of the other three groups. "Pretty" is used most heavily by Asians and Hispanics. Black speakers appear to be quite infrequent users of "really" and "pretty" as compared to the other three groups. Linear mixed effects models fit to the data find somewhat weak yet statistically significant effects of ethnicity with respect to frequencies of "really" ($\eta^2 = 0.029$, $p < 0.001$), "very" ($\eta^2 = 0.019$, $p < 0.05$), and "pretty" ($\eta^2 = 0.016$, $p < 0.05$).

4.2.4.4 Discourse markers

While most discourse markers appear in figure 4.30 to be used at roughly the same frequency among the various ethnic groups, a few discourse markers stand out as being preferred by one group or another. The markers "yeah" and "so" for instance are highly preferred by Asian speakers, while usage of "you know" is dominated by Black and Hispanic speakers. "Okay" appears to be favored by Black and Asian speakers. "Like" interestingly appears to exhibit a pattern of avoidance, with Black speakers using this discourse marker roughly 25% less frequently than the other three groups.

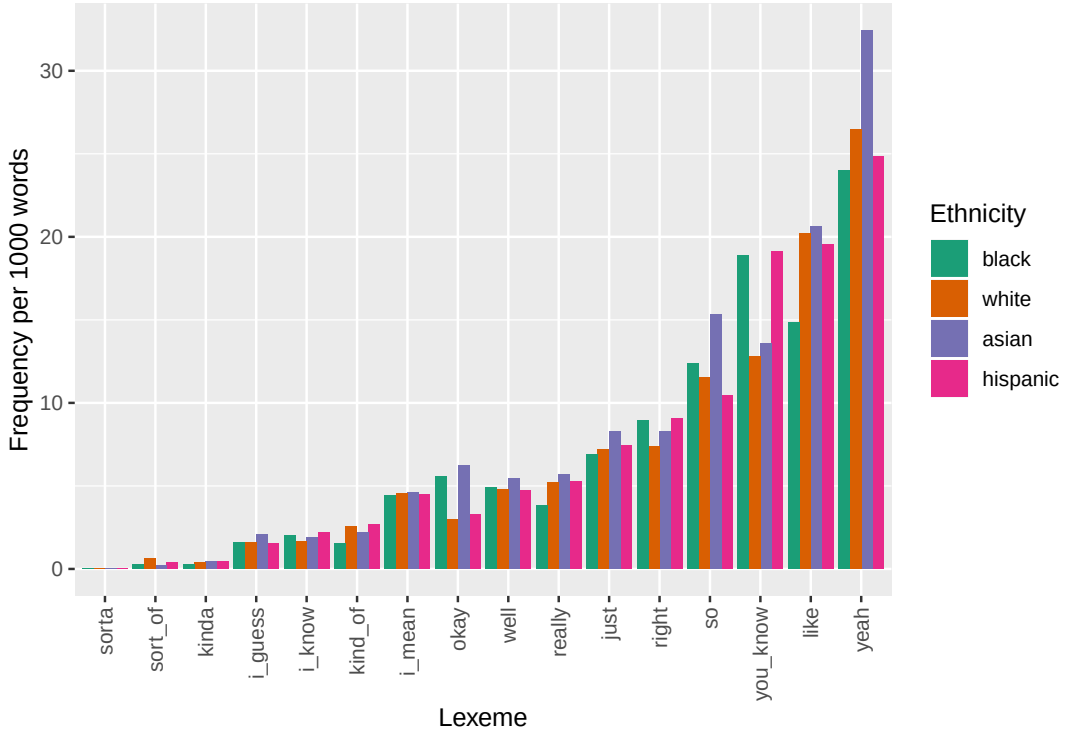

Figure 4.30: Discourse marker usage by ethnicity

Linear mixed effects models fit to the data reveal somewhat weak yet statistically significant effects of ethnicity on frequency of "really" ($\eta^2 = 0.021$, $p < 0.01$), "like" ($\eta^2 = 0.018$, $p < 0.05$), "okay" ($\eta^2 = 0.059$, $p < 0.001$), "yeah" ($\eta^2 = 0.015$, $p < 0.05$), "so" ($\eta^2 = 0.055$, $p < 0.001$), "sort of" ($\eta^2 = 0.017$, $p < 0.05$), and "you know" ($\eta^2 = 0.025$, $p < 0.01$).

4.2.4.5 Pronoun proportions

Figure 4.31 shows the proportion with which speakers from the four ethnic groups used first, second, and third person pronouns. While there is some variation, the differences appear minimal. Linear mixed effects models fit to the data do however show significant (albeit somewhat weak) effects of ethnicity for the first person ($\eta^2 = 0.034$, $p < 0.001$), second person ($\eta^2 = 0.027$, $p < 0.01$) and third person ($\eta^2 = 0.022$, $p < 0.01$) pronoun proportions.

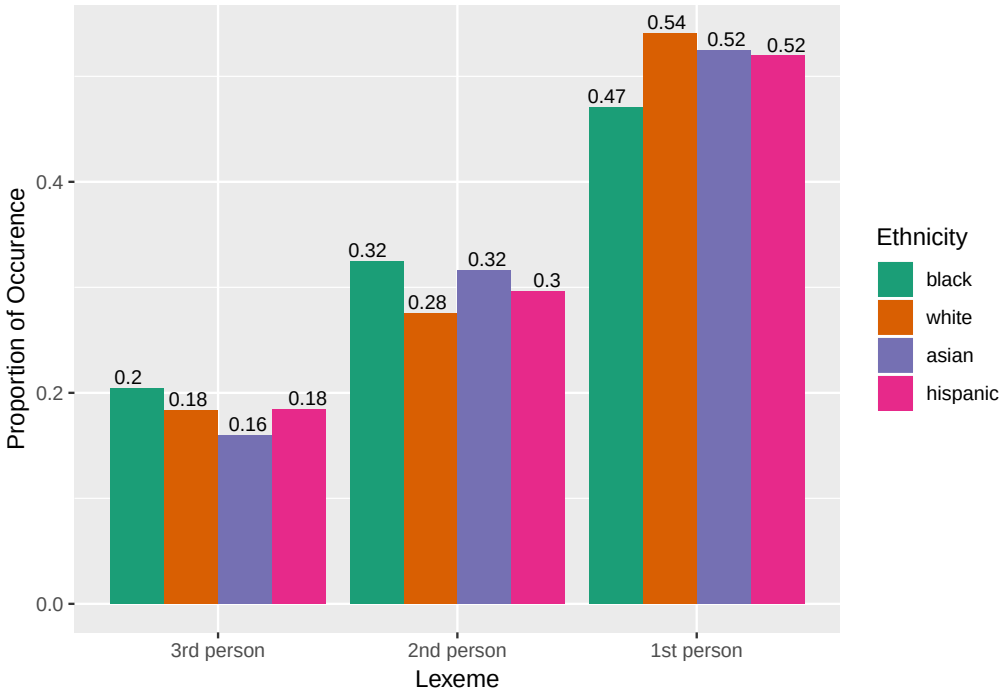

Figure 4.31: First, second, and third person pronoun usage proportions

4.2.4.6 Politeness and taboo frequency

Figure 4.32 shows the kword frequency for politeness and taboo-related terminology among the four ethnicities considered in this dissertation.

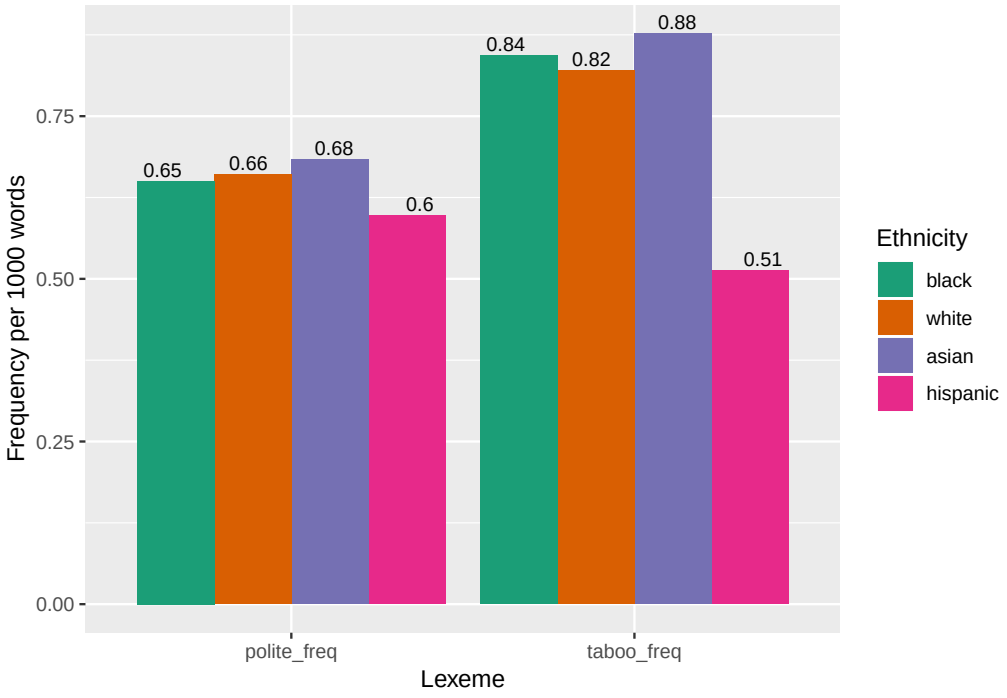

Figure 4.32: Usage of taboo and politeness terms by ethnicity

With regard to politeness terminology, Black, White, and Asian speakers all appear to exhibit roughly similar usage levels, with Hispanic speakers using politeness terminology somewhat less frequently than the other three groups.

For taboo word frequency, it appears that Asian speakers tend to use taboo words with the highest frequency, followed by Black and White speakers. Hispanics use taboo words at the lowest frequency of all ethnic groups considered– roughly 60% of the rates for the other three groups.

A linear mixed effects model found no significant effect of ethnicity on either of these features, indicating that the low values for Hispanic speakers may be an artifact of small sample size for the Hispanic speakers.

4.2.4.7 Sentiment

Figure 4.33 presents the subjectivity and polarity of speech segments in the corpus differentiated by ethnicity.

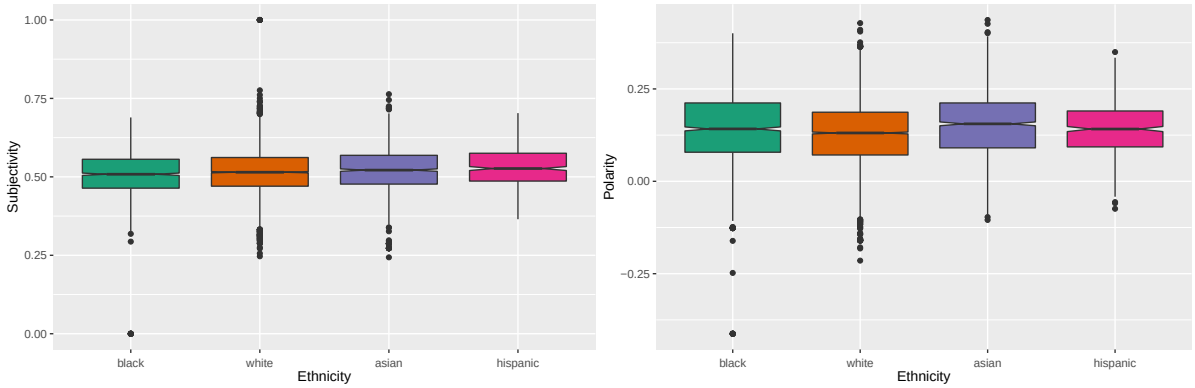

Figure 4.33: Segment subjectivity (left) and polarity (right) by ethnicity

There appears to be little difference in either of these metrics based on ethnicity, though it does appear that White speakers may exhibit lower polarity on average than the other groups while Asian speakers may exhibit higher polarity than the other groups. A linear mixed effects model fit to the polarity data found a weak yet significant effect of ethnicity on segment polarity ($\eta^2 = 0.018$, $p < 0.05$). No significant difference among the ethnicities however was found for subjectivity.

4.2.4.8 Speech rate and word length

Figure 4.34 presents a comparison of the speech rate (tokens per minute) and average word length (in syllables) between speakers of the four ethnic groups considered.

There appears to be somewhat of a two-way distinction among the four ethnicities when it comes to word length, with White and Asian speakers tending to produce words with slightly more syllables per word on average than Black and Hispanic speakers. A

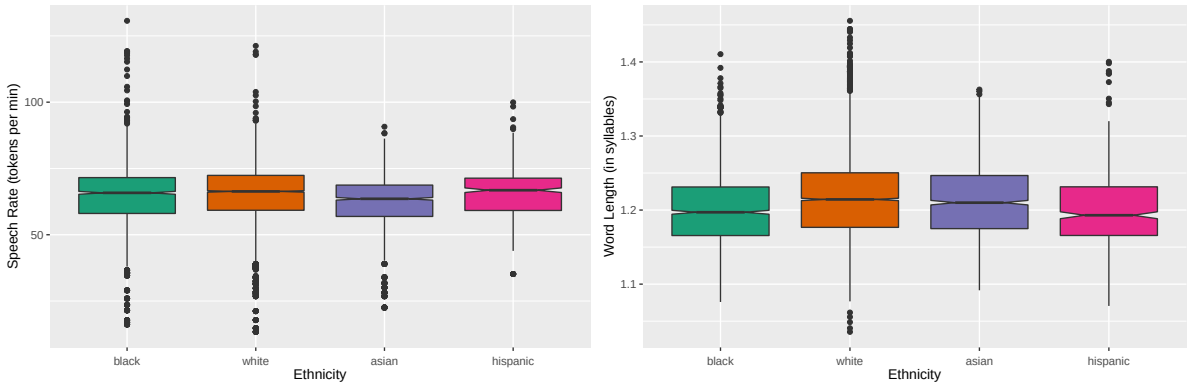

Figure 4.34: Speech rate (left) and word length (right) by ethnicity

linear mixed effects model found a weak yet significant effect of ethnicity with respect to this feature ($\eta^2 = 0.017$, $p < 0.05$).

With regard to speech rate, Black, White, and Hispanic speakers all appear to exhibit extremely similar speech rates, while Asian speakers appear to exhibit slightly slower speech rates than the other three groups. As with word length, a linear mixed effects model fit to the data finds a weak yet significant effect of ethnicity on this feature ($\eta^2 = 0.018$, $p < 0.05$).

4.2.4.9 Top informative ngrams

As detailed in chapter 3, the top 2000 ngrams (unigrams and bigrams combined) according to information gain for each sociodemographic trait of focus were extracted to serve as predictor features for the models presented in chapters 5 and 6. Table 4.2 presents the top 20 ngrams according to information gain for speaker ethnicity. The full 2000 are not shown here for reasons of space.

Table 4.2: Top 20 informative ngrams for ethnicity

attributes	infogain
uhhuh_oh	0.0285783
try	0.0274223
oh_you	0.0247547
okay_okay	0.0229029
school_oh	0.0223360
um_i	0.0215278
okay	0.0210200
okay_so	0.0204903
be_pretty	0.0204345
way_you	0.0194375
man_you	0.0192374
school_yeah	0.0189342
oh_oh	0.0185524
oh_okay	0.0180145
now_oh	0.0176348
as_far	0.0167823
hard_it	0.0165488
student	0.0162436
negative	0.0160057
domestic	0.0157866

4.3 Age

4.3.1 Overall age distribution

Before examining any particular predictor with respect to age, it's important to consider the overall distribution of age within the data-set. Figure 4.35 shows a histogram overlaid with a density plot of speaker age for the speech segments used in this dissertation. The corpus appears rather heavily skewed towards younger speakers, with very few speakers 60 years old or older. This may cause the models to over-predict younger age ranges, and will likely make it difficult to recognize those speakers 60+ years old. As with the class imbalance for ethnicity, this issue is addressed in chapter 5.

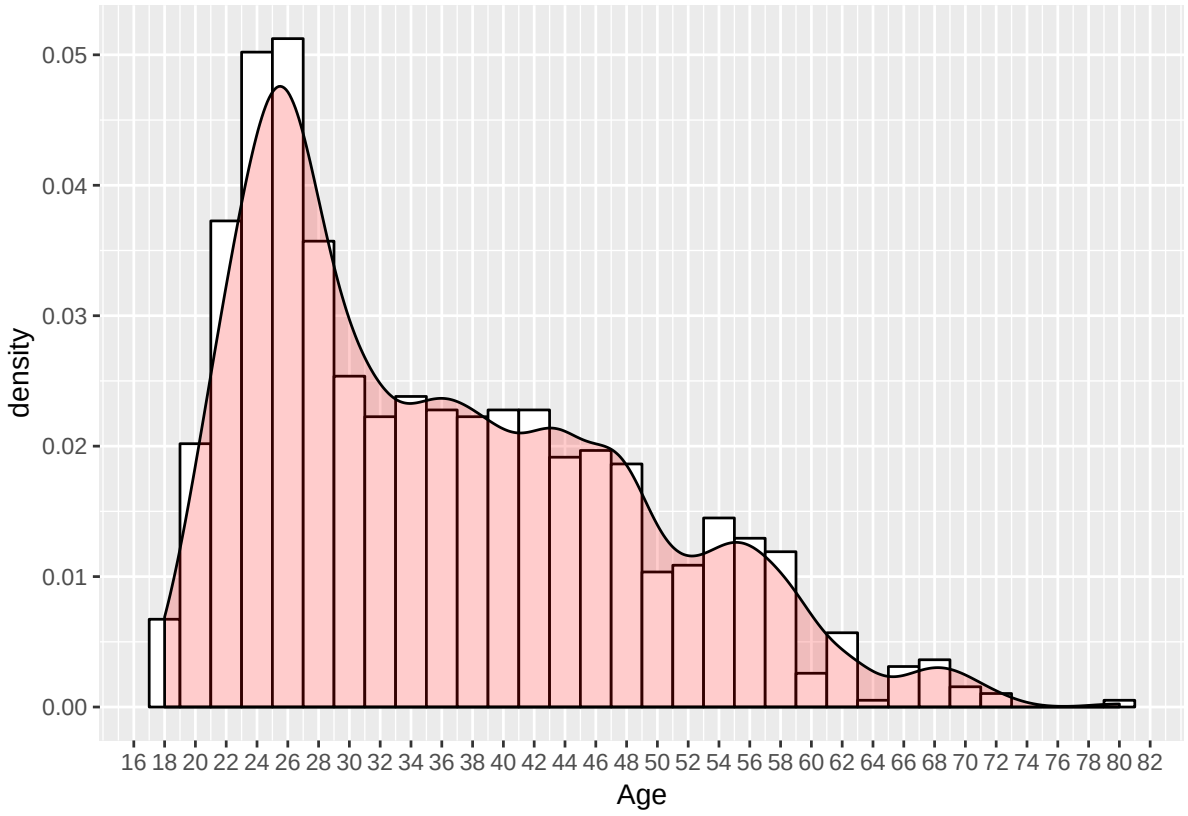

Figure 4.35: Overall age distribution

Though age may be treated as a categorical or continuous variable, the following analysis (as well as the rest of the discussion on age in the following chapters) treats age as a categorical variable as described in chapter 3 for the reasons laid out in chapters 2 and 3. For the purpose of this investigation, age is binned into the following five age range categories: ages 16-25, ages 26-35, ages 36-45 , ages 46-55 , and ages 56+.

4.3.2 Acoustic variables

4.3.2.1 HNR

As figure 4.36 shows, HNR values do not appear to show much significant movement across the age spectrum, although there does appear to be a slight bump in HNR for

speakers in the older two age groups. A linear mixed effects model fit to the data does not show any significant effect of age category on HNR values.

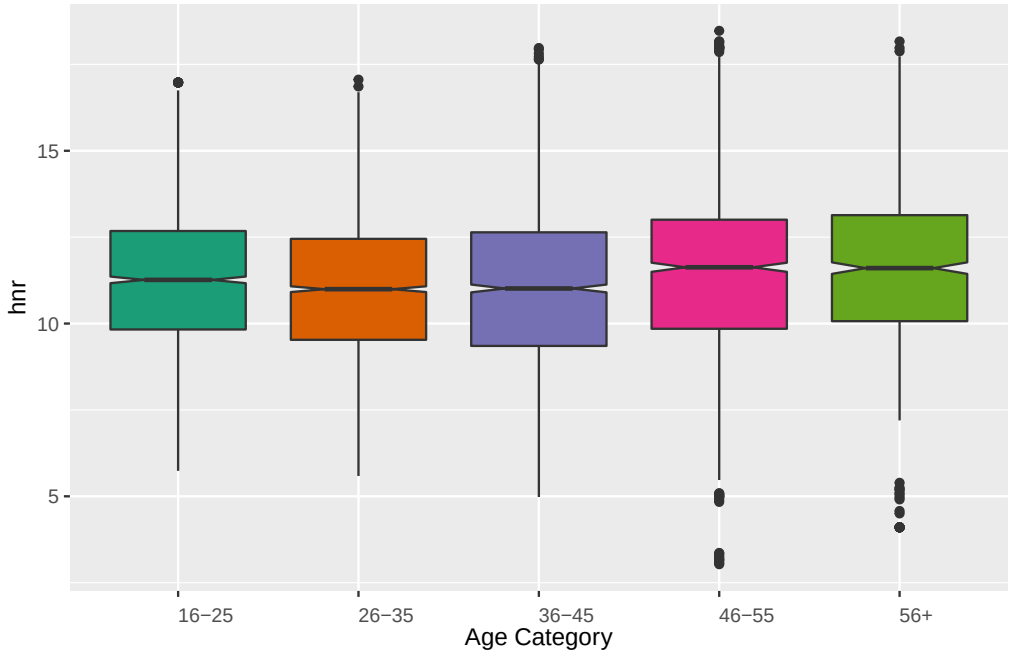

Figure 4.36: HNR measurements by age

4.3.2.2 Jitter

Though the RAP and PPQ5 measurements of jitter don't appear to show much variation across age groups, absolute jitter appears to exhibit a pattern of steadily increasing jitter values up until the oldest age group, in which it declines. Absolute jitter may therefore be a useful indicator for distinguishing between speakers in the middle three age categories. Linear mixed effects models fit to the data found a weak yet statistically significant effect of age category on absolute values of jitter ($\eta^2 = 0.017$, $p < 0.05$) but not on the other two jitter measurements.

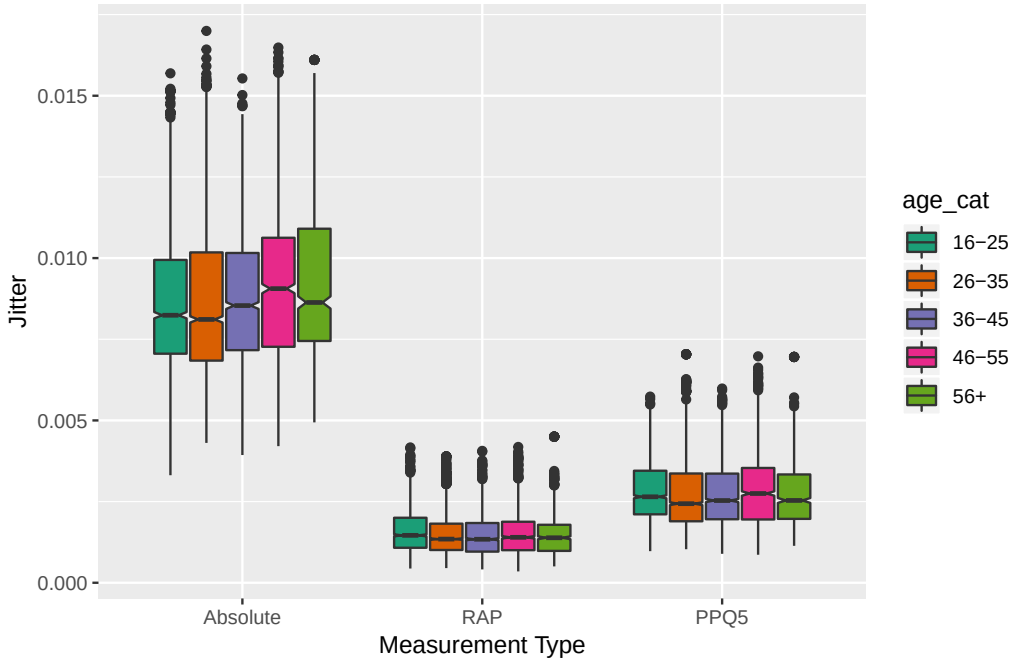

Figure 4.37: Jitter measures by age

4.3.2.3 Shimmer

Unlike jitter and HNR, absolute shimmer appears to show a slight but steady downward trend until around age 55, at which point it levels out. Though this pattern is more profound in the absolute measure of shimmer versus the APQ3 measure, the same basic trend is observed for both, indicating that shimmer may be a useful indicator for distinguishing between age categories. Linear mixed effects models show moderately weak yet statistically significant effects of age group on both the absolute ($\eta^2 = 0.041$, $p < 0.001$) and APQ3 ($\eta^2 = 0.034$, $p < 0.001$) measures of shimmer.

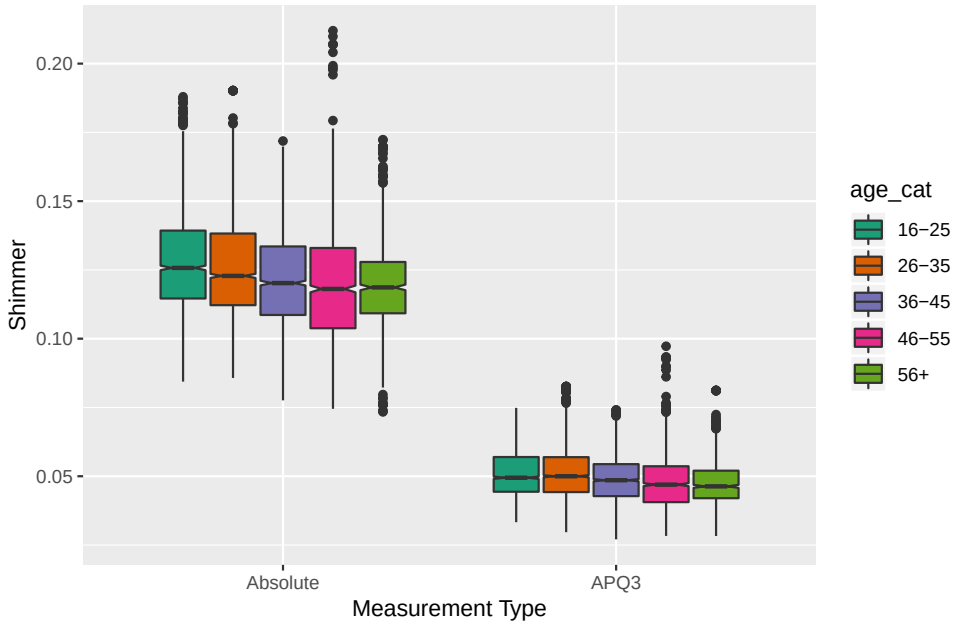

Figure 4.38: Age category differences in shimmer at the segment level

4.3.2.4 Pitch

Pitch minimum and mean values appear relatively flat throughout the age groups. However, there does appear to be somewhat of a downward trend in maximum pitch starting from the youngest age group with the highest values to the middle age group, after which max pitch appears to hold steady. Linear mixed effects models fit to the data show relatively weak yet statistically significant effects of age category on measures of maximum pitch ($\eta^2 = 0.026$, $p < 0.01$), but not minimum or mean pitch.

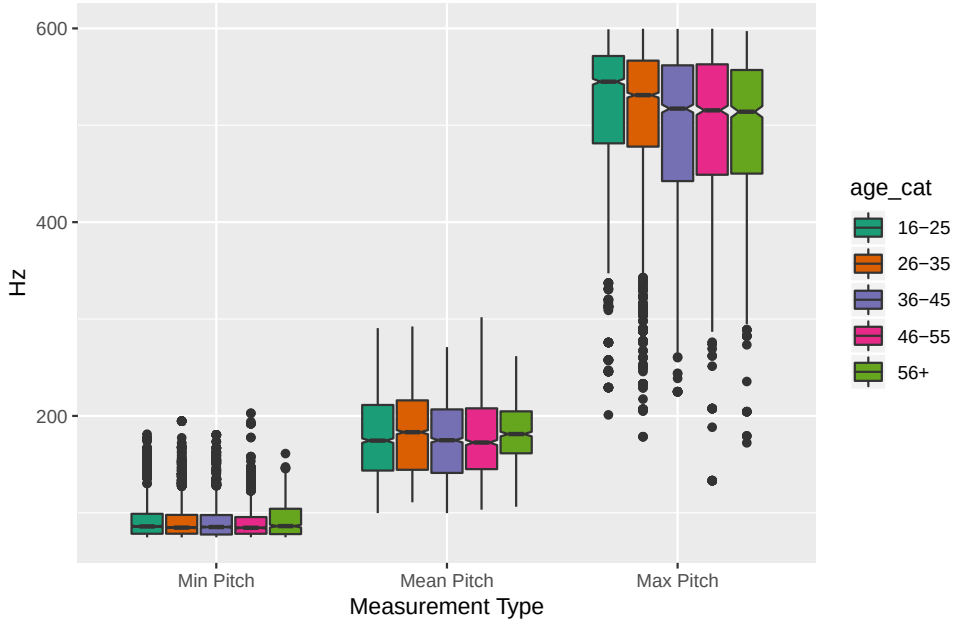

Figure 4.39: Age category differences in pitch at the segment level

4.3.3 Phonetic features

4.3.3.1 Vowel space

Vowel space area and dispersion appear from figures 4.40 and 4.41 to exhibit a growth and decline trend wherein mean values increase from the youngest to the middle-most age categories and subsequently decrease from the middle-most to the oldest categories. Linear mixed effects models fit to the data confirm relatively weak yet statistically significant effects of age category on both vowel space area ($\eta^2 = 0.015$, $p < 0.05$) and vowel dispersion ($\eta^2 = 0.016$, $p < 0.05$). Why these measures might exhibit this "rise-and-fall" pattern across the age groups is unclear, as I am aware of no existing work demonstrating such a trend. One possibility is that the middle age groups may be under more pressure to produce "clear speech" (see e.g. Ferguson and Kewley-Port, 2007) than other age groups on average as a result of caregiver and/or marketplace-induced pressures (e.g. modeling speech norms for children, sounding "professional" in

the workplace, etc.).

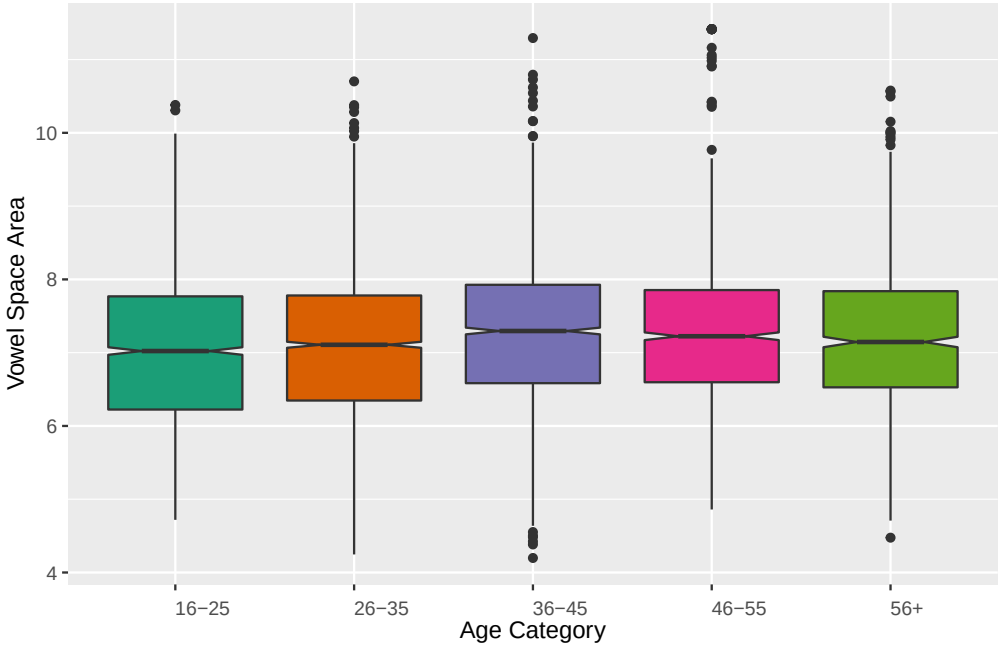

Figure 4.40: Age differences in vowel space area

Unlike vowel space area and dispersion, figure 4.42 shows no meaningful difference in terms of vowel dynamicity across age groups, as was expected.

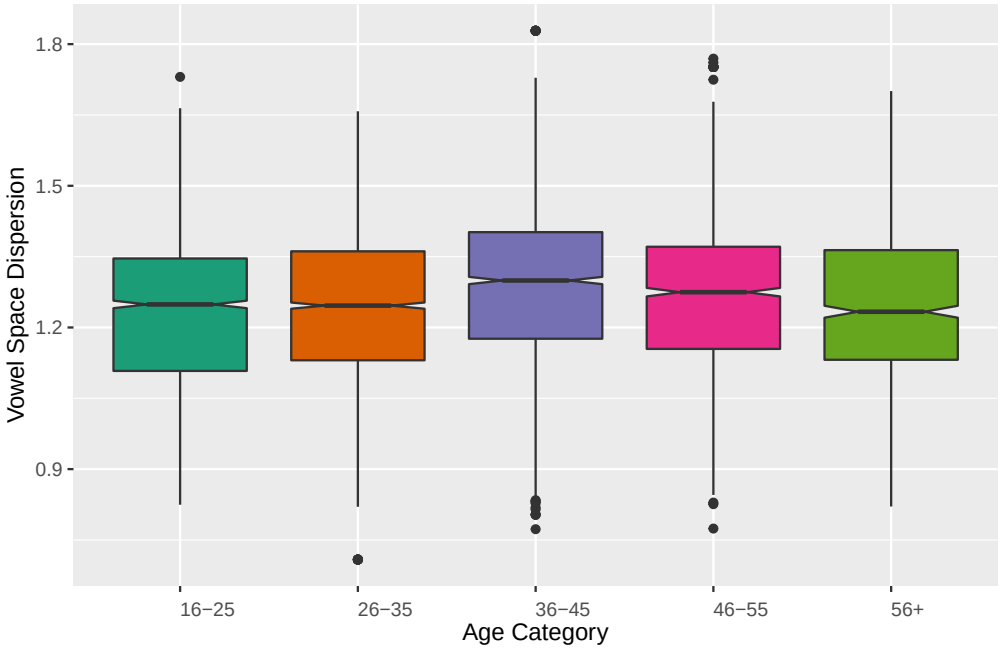

Figure 4.41: Age differences in vowel space dispersion

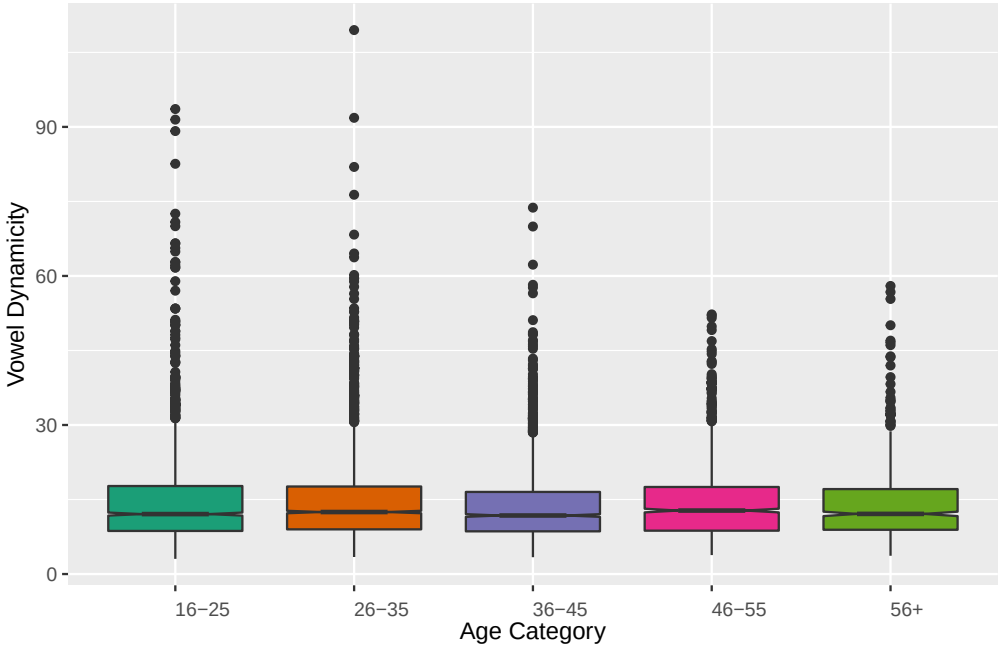

Figure 4.42: Age differences in vowel dynamicity

4.3.3.2 Vowel trajectories

Figure 4.43 plots the vowel trajectories for all five age groups in accordance with the plotting conventions laid out above in section 4.1.3.2.

The salient differences between age groups apparent from figure 4.43 are discussed in detail below. Refer to appendix C for a list of which specific vowels at which specific points show a significant effect of age.

- **EY**: Speakers in the 16-25 age group and the 26-35 age group appear to realize EY with a shorter overall trajectory and a relatively fronted and raised onset than speakers in the older age groups.
- **EH**: The onset of EH appears slightly lowered in the 16-25 and 26-30 age groups as compared to the older age groups.
- **AE**: Speakers in the youngest age group (16-25) exhibit a relatively lowered onset of AE as compared to the older groups.
- **UH**: UH appears to be more fronted in the younger age groups of this data-set. Speakers in the 56+ age group produce UH with an relatively backed onset compared to speakers in the 26-35, 36-45, and 46-55 age groups, and speakers in age groups 26-35, 36-45, and 46-55 in turn all appear to produce UH with a relatively backed onset compared to speakers in the 16-25 group.
- **OW**: The trajectory of OW appears to be shorter for speakers in the younger age groups as compared to the older age groups. While all groups have relatively similar onset positions for OW, speakers ages 36+ appear to exhibit a more diphthongal, backed offglide as compared to speakers in the 16-25 and 26-35 age groups.
- **AA**: Likewise AA for speakers ages 36 and above appears more diphthongal than younger speakers, with clear movement between the midpoint of the vowel and the off-glide measurement point. Speakers in age groups 16-25 and 26-35 in comparison show almost no movement whatsoever between the midpoint and offglide measurement points.
- **AY**: There appears to be a clear difference in realization of AY between age groups 36+ and age groups younger than 36. Speakers in the older groups exhibit onsets and off-glides for AY that are relatively retracted and lowered compared to speakers in the younger age groups.
- **AW**: Speakers in age groups 36+ exhibit relatively fronted onsets and offglides for AW than younger speakers.

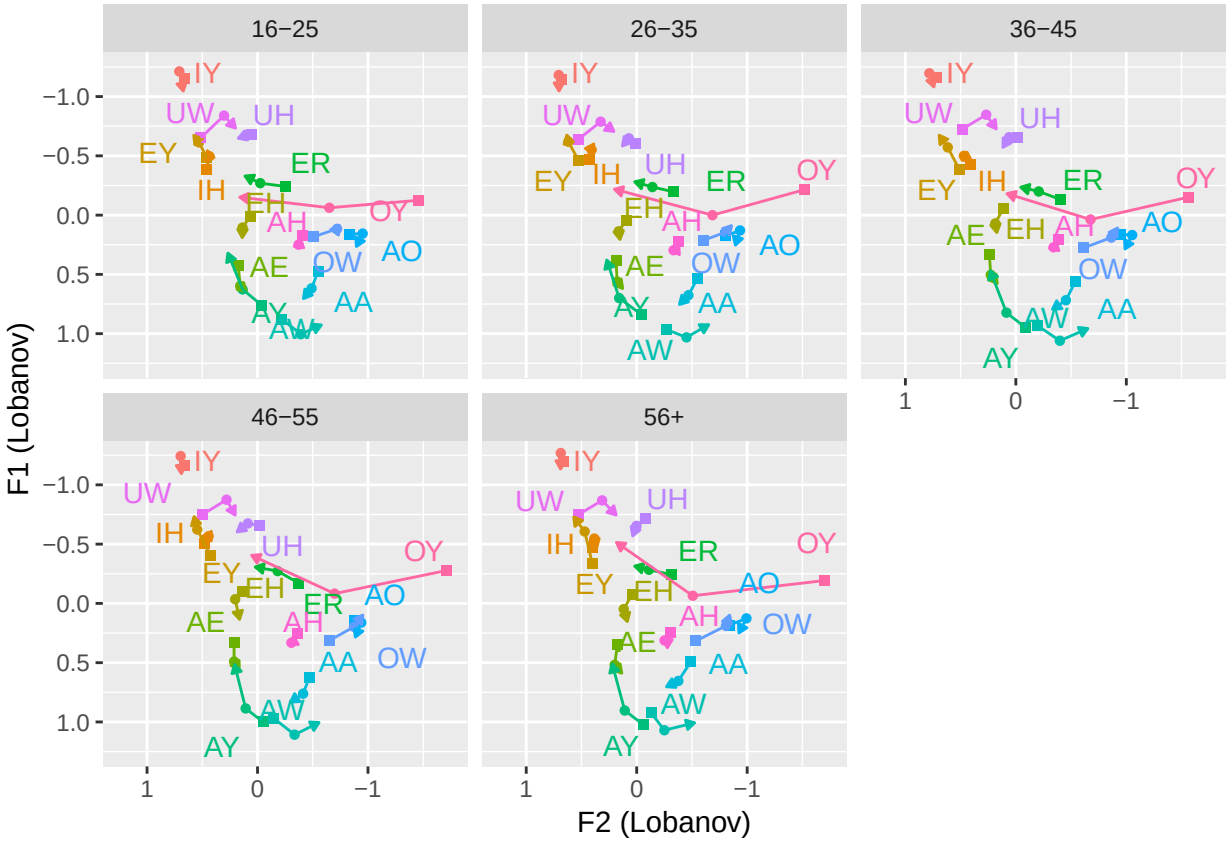

Figure 4.43: Vowel trajectories by age group

4.3.4 Lexical features

4.3.4.1 Quotatives

Figure 4.44 displays the average kword frequency of quotatives for each of the five age groups.

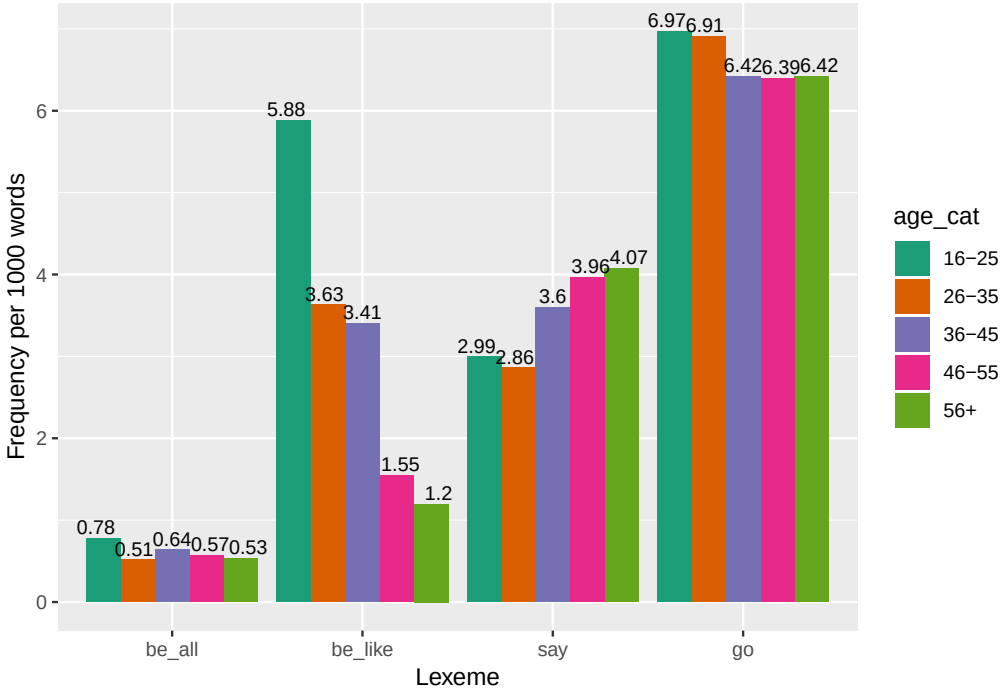

Figure 4.44: Quotative usage by age

The most striking and salient difference among the age groups is the frequency with which the "be like" construction was used. There appears to be a strong trend wherein younger speakers use this construction far more frequently than older speakers. A relatively strong, significant effect of age group on rates of "be like" is confirmed by a linear mixed effects model ($\eta^2 = 0.137$, $p < 0.001$).

Similarly, there appears to be an inverse (though somewhat weaker) trend for "say," with younger age groups using this quotative successively less frequently than older age groups. Despite the visual trend however, a linear mixed effects model fit to

the data does not show a significant effect of age on frequency of "say."

The trends observed here are in line with the research discussed in chapter 2, which suggest that usage of the "be like" quotative construction has been gaining ground relative to usage of more traditional quotatives like "say" within the last several decades.

4.3.4.2 Modals

Figure 4.45 shows the kword frequency of modal and semi-modal constructions across the five age groups.

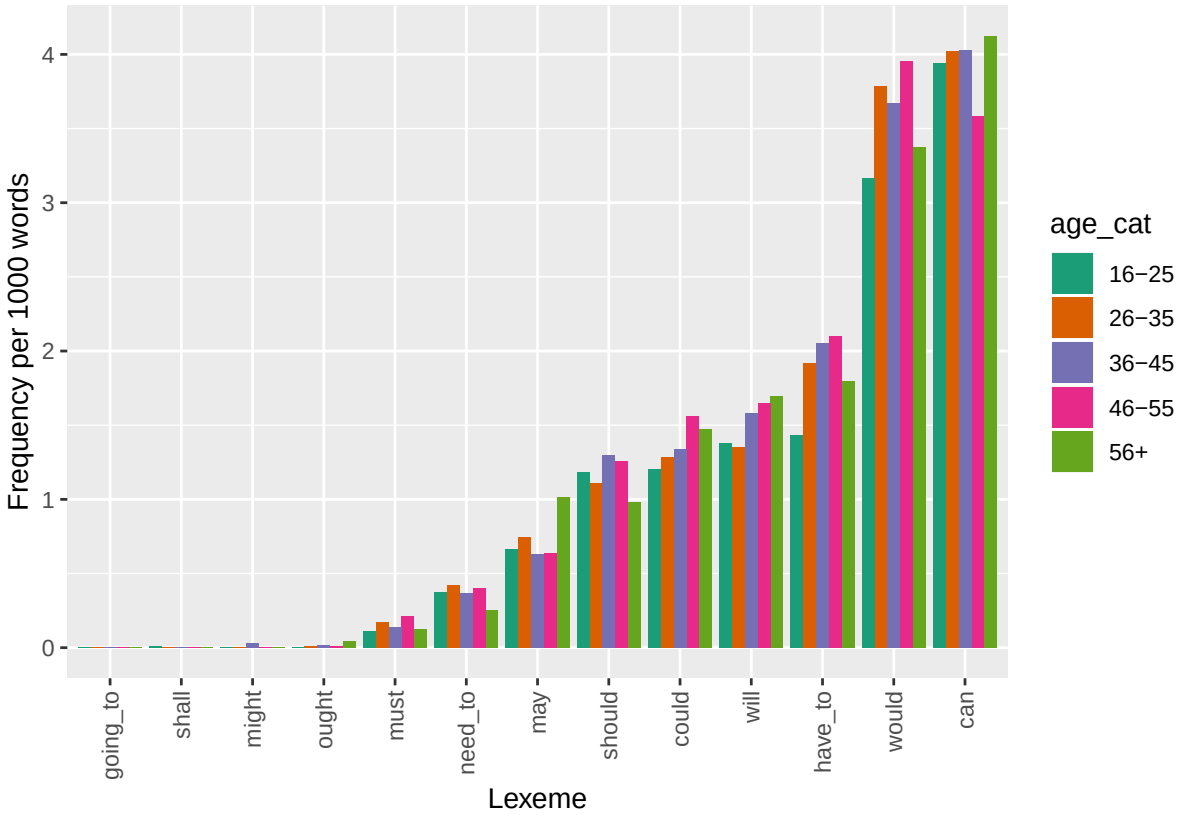

Figure 4.45: Modal usage by age

Though there are no clear linear trends as there are with quotatives, some interesting differences do emerge. Recall from chapter 2 that Barbieri (2008) observed significantly higher usage of "may," "will," "could," "ought," "might," and "have to"

for older speakers as compared to younger speakers. In most cases, the data from the NIST corpus used here are in line with those findings. Barbieri also found however that the one modal verb used more by younger speakers than older speakers was "can." This does not appear to be the case in the NIST data, as the oldest age group exhibits roughly the same frequency for "can" as the youngest three age groups.

Despite the apparent differences in figure 4.45, the only statistically significant difference between age groups with respect to modal frequency found during linear mixed effects testing was for the modal construction "have to" ($\eta^2 = 0.015$, $p < 0.05$).

4.3.4.3 Intensifiers

Figure 4.46 compares the kword frequency of select intensifier usage across the five age groups.

There appears to be clear differentiation among the age groups with respect to usage of most of the intensifiers considered here, with several exhibiting the sort of linear trend that one would expect from a variable that is steadily shifting over time. Perhaps the clearest difference of all is that between usage of "very" for the oldest group as compared to the younger four groups. The difference in usage of "very" among age groups was found to be statistically significant in a linear mixed effects model ($\eta^2 = 0.016$, $p < 0.05$).

"Really" appears to show a two way distinction, with speakers in the youngest two age groups producing this intensifier nearly twice as frequently as speakers in the older three groups. A linear mixed effects model confirms a weak yet significant effect of age on usage of "really" as an intensifier ($\eta^2 = 0.039$, $p < 0.001$).

"Pretty" shows the clearest linear trend, with usage steadily increasing across age

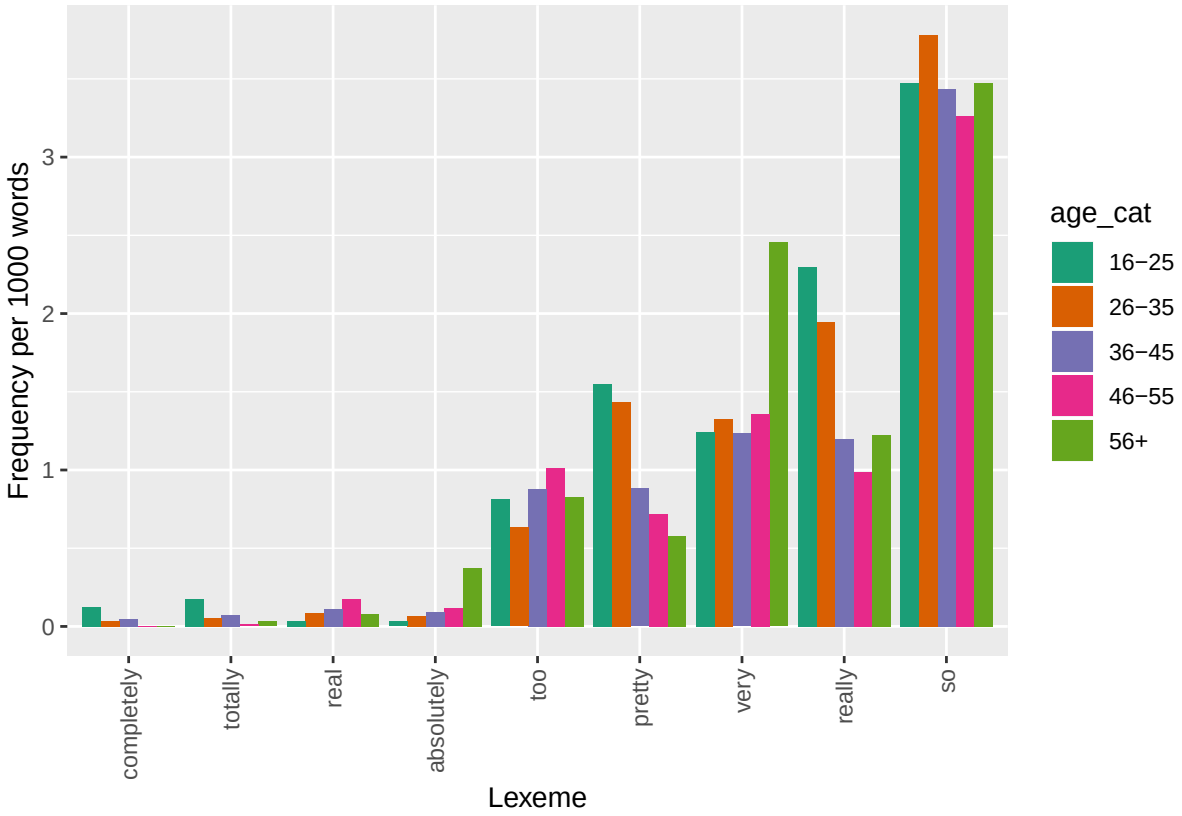

Figure 4.46: Intensifier usage by age

groups from older to younger. As with "really," a linear mixed effects model confirms a weak yet significant effect of age on usage of "pretty" as an intensifier ($\eta^2 = 0.035$, $p < 0.001$).

Though much less frequent overall than "pretty," the intensifier "absolutely" also shows a somewhat linear (though inverse) usage trend, with usage decreasing over age groups from older to younger. Though weak, a linear mixed effects model fit to the data shows a significant effect of age on usage of this intensifier as well ($\eta^2 = 0.018$, $p < 0.05$).

The two intensifiers "totally," and "completely" are used rarely if at all by speakers in the two oldest age groups, and are used most by speakers in the youngest age group.

All these findings are in line with Barbieri (2008), who found young speakers in

the BNC leading in usage of most intensifiers with the exception of a few particular intensifiers such as “very” and “absolutely.”

4.3.4.4 Discourse markers

Figure 4.47 shows a comparison of the kword frequency for select discourse markers across the five age groups considered here.

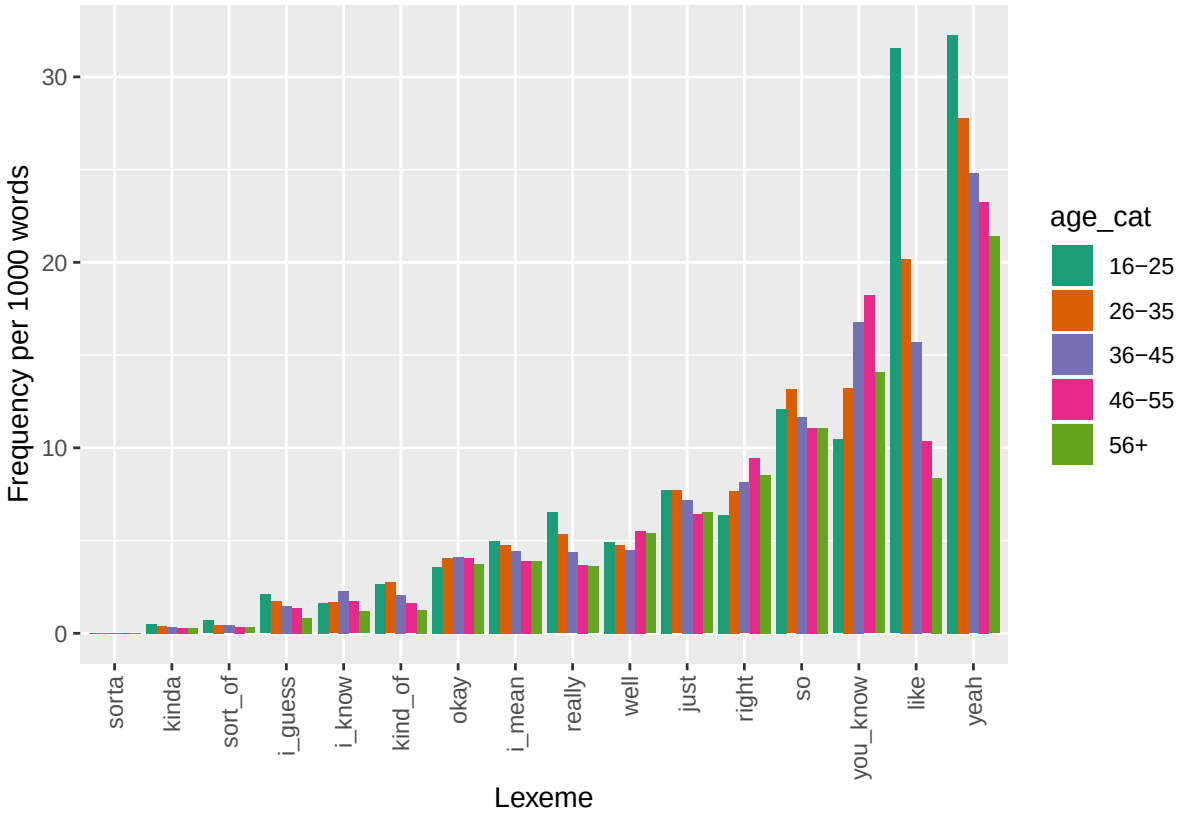

Figure 4.47: Discourse marker usage by age

As expected, nearly all of the discourse markers under consideration show signs of either increasing or decreasing from younger to older age groups, making the kword frequency of these lexemes likely highly informative to models trained to predict speaker age. The two most frequent discourse markers overall– “yeah” and “like”– also appear to show the sharpest age group distinction, with younger speakers using these markers

at much higher rates than older speakers. The markers "so," "just," "really," "I mean," "I guess," "kinda/kind of," and "sorta/sort of" also appear to be higher in frequency for younger speakers, albeit less clearly delineated than the differences for "like" and "yeah." In contrast, the markers "you know" and "right" appear to be lower for speakers in the younger age groups. Though significant effects of age were found for several of the discourse markers examined here (refer to appendix C for a full list), a particularly strong effect of age was found for the discourse marker like ($\eta^2 = 0.239$, $p < 0.001$).

4.3.4.5 Pronoun distribution

Figure 4.48 shows the relative proportion with which each age group used first, second, and third person pronouns.

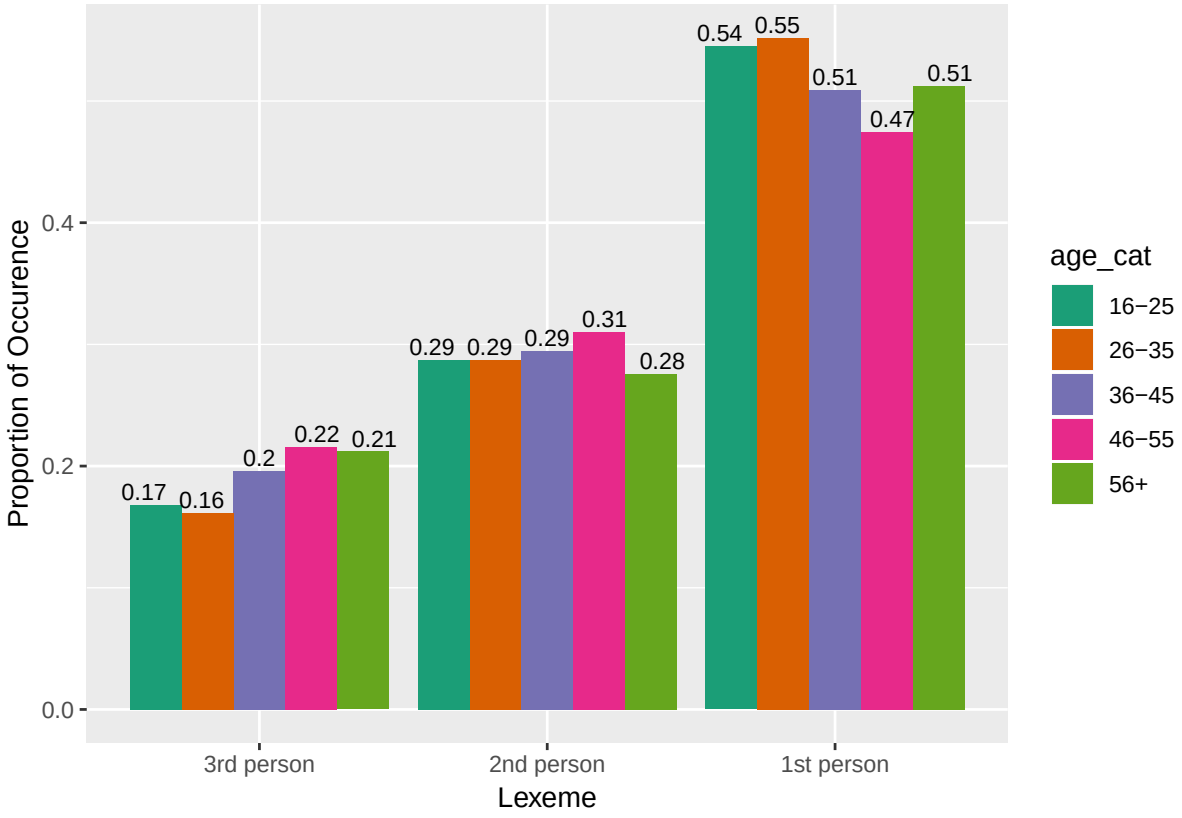

Figure 4.48: First, second, and third person pronoun usage proportions

Though the differences are somewhat small, it does appear that the younger two age groups use first person pronouns at a higher rate than the other three groups, and vice versa for third person pronouns. These differences are confirmed as statistically meaningful in a linear mixed effects model (first person pronouns: $\eta^2 = 0.043$, $p < 0.001$; third person pronouns: $\eta^2 = 0.046$, $p < 0.001$), and are in line with the findings from Barbieri (2008) that younger speakers in the BNC tend to use first person pronouns more frequently and third person pronouns less frequently than older speakers.

4.3.4.6 Taboo and politeness terminology

Figure 4.49 compares the kword frequency of taboo and politeness terminology across the five age groups.

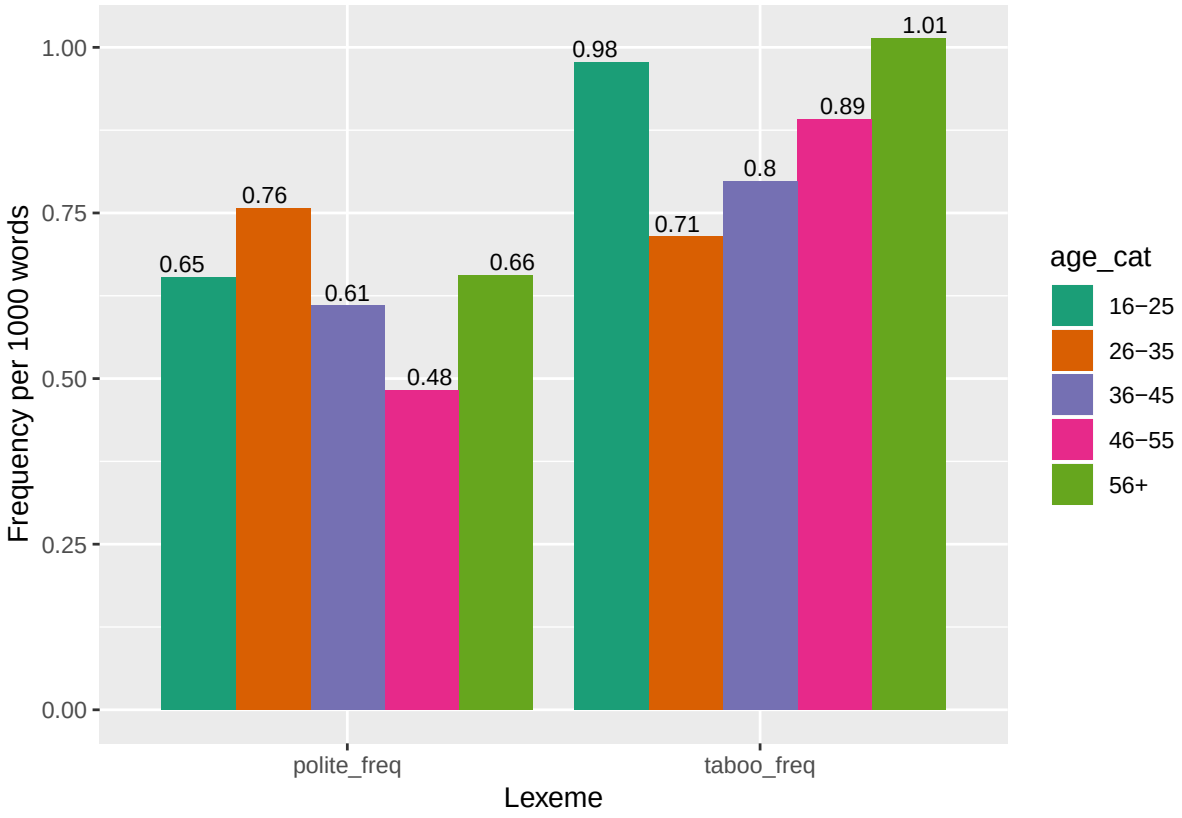

Figure 4.49: Usage of taboo and politeness terms by age

Focusing first on the usage difference for taboo terminology, we see the classic "u-shaped" pattern of usage often seen with non-standard, age-graded variables. As mentioned in chapter 2, this u-shaped usage pattern seen for non-standard variables is characterized by a peak in adolescence when pressure to not conform to societal norms is highest, A trough for working-age speakers, and a second peak in older age as speakers leave the workforce and pressure to conform to societal norms is somewhat relaxed. The usage of taboo terminology in the NIST corpus appears to follow this pattern exactly, with peaks in the oldest and youngest age groups, and a trough for the middle three age groups.

Focusing now on the frequency of politeness terminology, we also see somewhat of a u-shaped pattern here, though this time the second youngest and not the youngest age group exhibits the initial peak. The difference between speakers 16-35 and speakers 36-55 is in line with the finding from Barbieri (2008) that younger speakers tended in the BNC to use "polite speech-act formulae" more frequently than older speakers. The peak in politeness terminology for older speakers however is not something she or others have observed, and may be an artifact of the small sample size of speakers 56+ years of age.

Despite the observed pattern of age-related differences for these two predictors, neither mixed effects model for these features reveals particularly strong or significant age-group effects.

4.3.4.7 Sentiment

Figure 4.50 compares polarity and subjectivity across the five age groups.

As this figure shows, there is essentially no difference in polarity whatsoever among

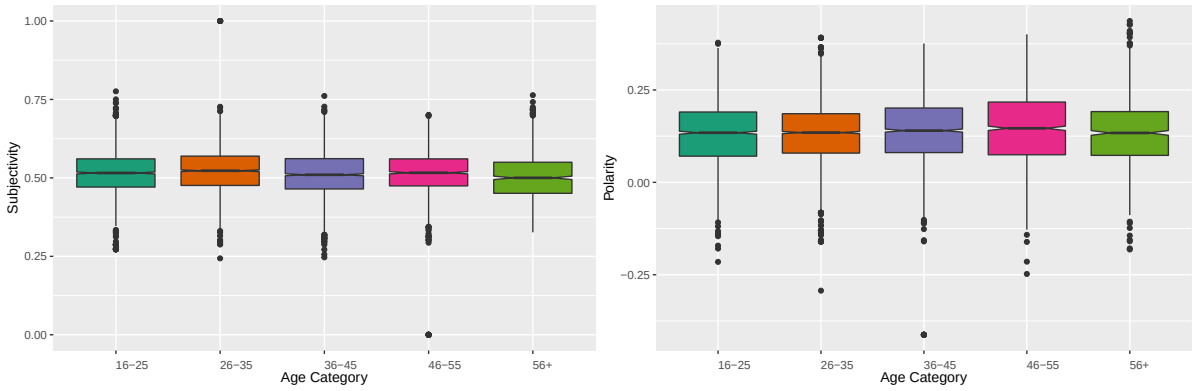

Figure 4.50: Segment subjectivity (left) and polarity (right) by age

the five age groups. Subjectivity also appears to show very little age-based variation, though there may be a very slight tendency for older speakers to be less subjective on average than younger speakers. Linear mixed effects models find no significant differences with respect to age for either of these two predictors.

4.3.4.8 Speech rate and word length

Figure 4.51 compares speaking rate and word length across the five age groups.

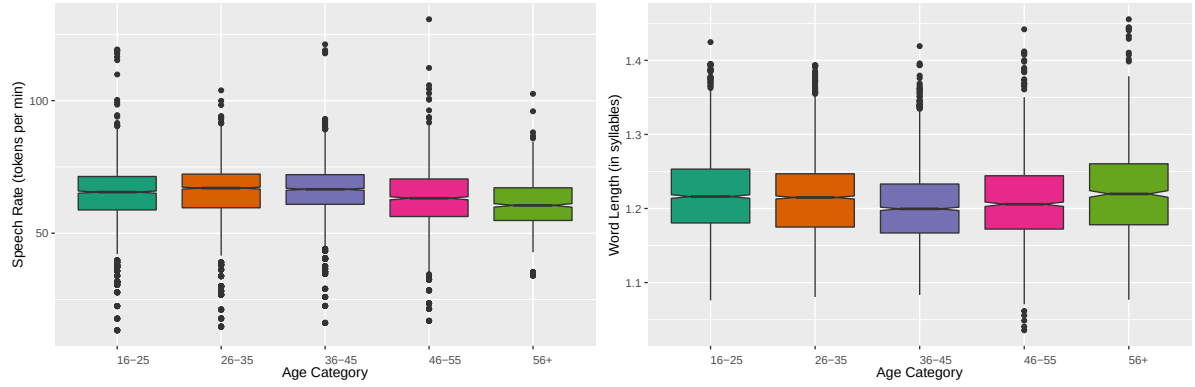

Figure 4.51: Speech rate (left) and word length (right) by age

Speaking rate appears to hold relatively constant for the three youngest age groups before exhibiting a slight decline across the oldest two age groups. The difference in speech rate between the oldest speakers and the speakers ages 16-45 may be a useful

feature in distinguishing those speakers belonging to the oldest age category. A linear mixed effects model fit to the data for speech rate finds a weak yet statistically significant effect of age ($\eta^2 = 0.016$, $p < 0.05$)

In contrast, word length appears to be highest on average for the oldest and youngest speaker groups, dipping among speakers in the 36-45 age category. As with speech rate, a linear mixed effects model fit to the data for word length finds a weak yet statistically significant effect of age ($\eta^2 = 0.017$, $p < 0.05$).

4.3.4.9 Top informative ngrams

As detailed in chapter 3, the top 2000 ngrams (unigrams and bigrams combined) according to information gain for each sociodemographic trait of focus were extracted to serve as predictor features for the models presented in chapters 5 and 6.

Table 4.3 presents the top 20 ngrams according to information gain for speaker age. The full 2000 are not shown here for reasons of space.

Given the striking age differentiation seen in usage of the quotative construction "be like" earlier in this section, it is not surprising (and is somewhat comforting) to see this construction ranked highest of all candidate ngrams in terms of information gain for the social trait of age. Likewise, the frequent presence of the discourse marker "like" in the bigrams making it into the top 20 for age echos the strong pattern of age differentiation for this discourse marker presented above. Presence of the slang term "cool" and the relational term "daughter" in ngrams making it to the top 20 for age also makes sense (e.g. the bigram "my_daughter" is likely used only by speakers who in fact have a daughter. The probability of a speaker having children may safely be assumed to increase with age, and therefore it makes sense that a speaker using

Table 4.3: Top 20 informative ngrams for age

attributes	infogain
be_like	0.0465302
cool	0.0445181
like_it	0.0404501
think_like	0.0367275
daughter	0.0330492
like_um	0.0315571
of_like	0.0309710
like_i	0.0288957
just_like	0.0287896
wife	0.0274536
of	0.0272810
feel_like	0.0271774
like_you	0.0264094
yeah_um	0.0252525
kind_of	0.0250567
completely	0.0241243
you_like	0.0235144
my_daughter	0.0234517
be_cool	0.0231503
say_they	0.0231163

this construction likely does not belong to the youngest age group.), and serves to highlight the importance of including the top n infogain ngrams as features rather than relying solely on lexical features which have been previously shown in the sociolinguistic literature to exhibit social patterning.

4.4 Region

4.4.1 General breakdown

Figure 4.52 shows the distribution of both speakers and segments for the four region groups considered in this dissertation.

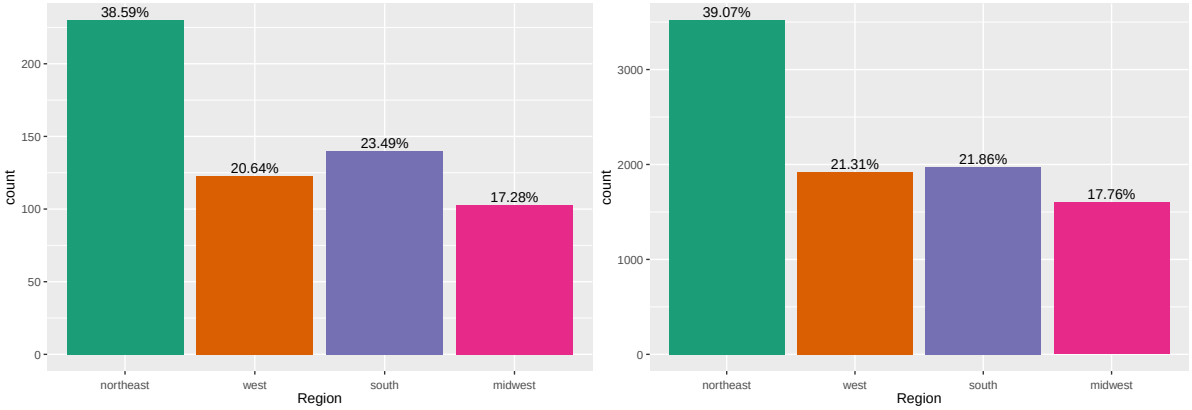

Figure 4.52: Breakdown of region by speakers (left) and segments (right)

There is somewhat of a class imbalance in the corpus among the four regions, with roughly twice the number of northeastern speakers as speakers from any other region. This class imbalance could be problematic down the line, and may lead to over-prediction of the "northeast" category. Strategies to rectify this class imbalance are discussed in more detail in chapter 5.

The proportion of segments from speakers in each region category is nearly identical to the regional proportion among the total number of speakers, indicating that speakers from the four regional groups took part in telephone conversations during data collection at roughly similar rates.

4.4.2 Acoustic features

4.4.2.1 Harmonic to noise ratio

Figure 4.53 compares the four regions examined here with respect to HNR.

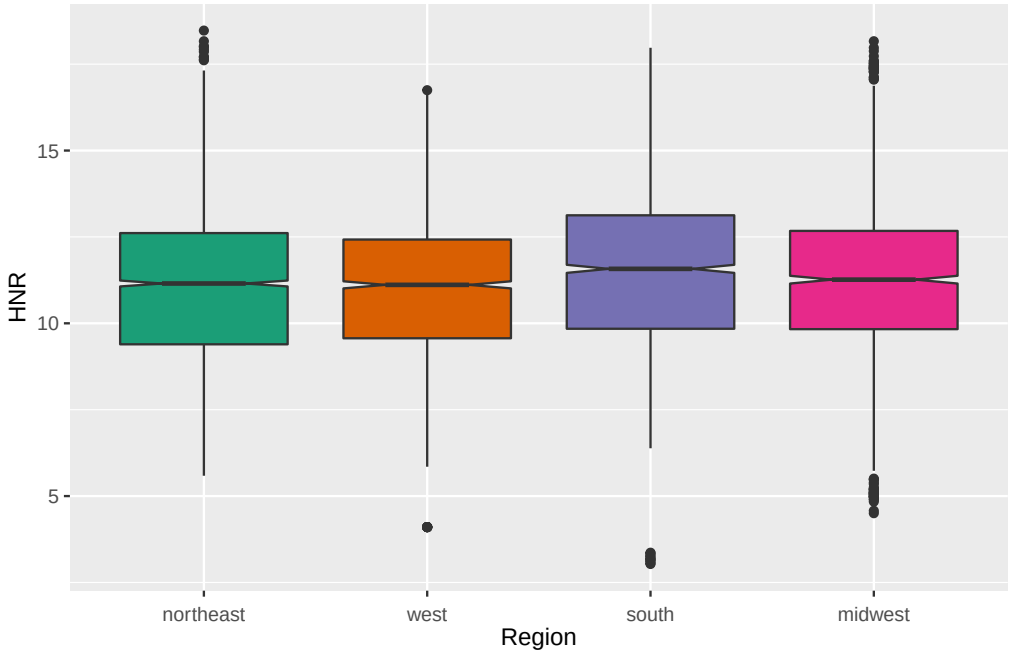

Figure 4.53: Difference in HNR between regions

Though the differences among regions are somewhat slight, figure 4.53 suggests that southern speakers may have somewhat higher values for HNR than the other three groups on average. That southern speakers would exhibit higher values for HNR than the other three regions would be somewhat surprising, as there is nothing in the literature to my knowledge that suggests this sort of effect. However, a linear mixed effects model found no significant effect of region on HNR values, indicating that the variation seen in figure 4.53 is not statistically meaningful.

4.4.2.2 Jitter

Figure 4.54 compares the four regions examined here with respect to three measurements of Jitter.

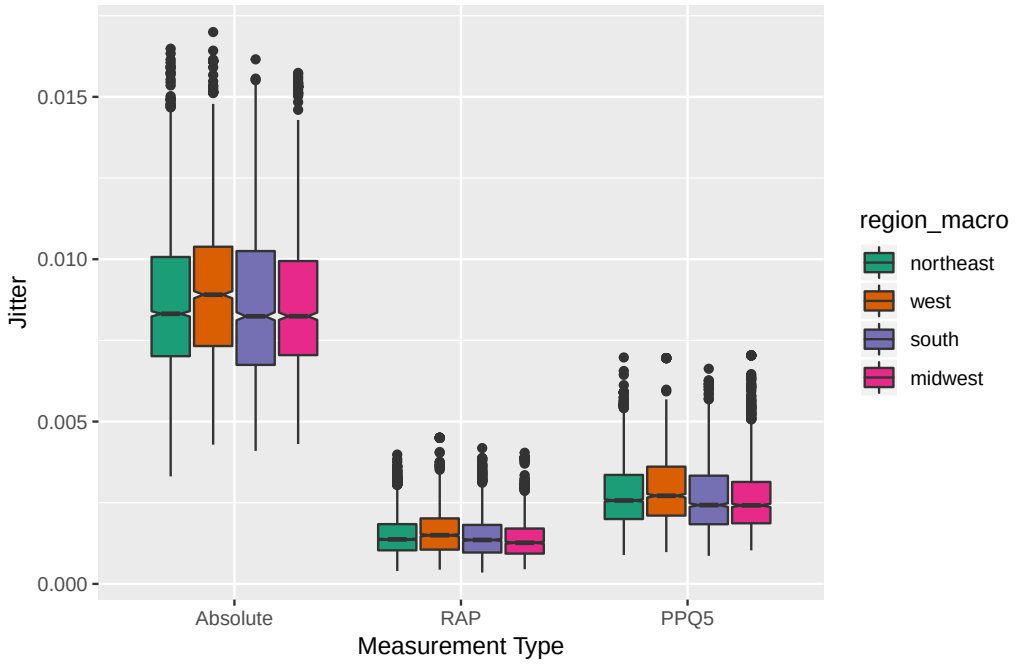

Figure 4.54: Region differences in jitter measurements at the segment level

Figure 4.54 appears to show some meaningful differences in jitter with respect to region. Speakers from the west appear to exhibit higher values for jitter on average than the other three groups, and speakers from the midwest appear to exhibit the lowest values for jitter. Speakers from the northeast and the south exhibit similar jitter values. A linear mixed effects model fit to the data confirms a significant difference between region groups for the RAP ($\eta^2 = 0.016$, $p < 0.05$) and PPQ5 ($\eta^2 = 0.014$, $p < 0.05$) measurements.

That jitter would exhibit regional differences is unexpected, as I am unaware of any claim in the literature to this effect. It's possible that the significant differences in jitter seen here are an artifact of an imbalance within the regional groups with respect

Table 4.4: Sex category percentages by region category

Sex	northeast	west	south	midwest
female	0.51	0.5	0.62	0.71
male	0.49	0.5	0.38	0.29

to some other confounding variable. The only other sociodemographic trait found to have a significant impact on jitter was speaker sex, with females exhibiting significantly lower values of jitter than males. Table 4.4 presents the percentage of male and female speakers in each regional group.

Speakers in the midwest group exhibit by far the highest percentage of female speakers. It seems likely therefore that the relatively low values for jitter in the midwest group may be ascribed to an imbalance with respect to sex. However, both the northeast and the west are split almost exactly evenly between male and female speakers, so it does not appear that sex is the driving factor behind the relatively high jitter values for speakers in the west as compared to the northeast and south.

4.4.2.3 Shimmer

There appears to be little meaningful difference if any between the four regions in terms of shimmer. A linear mixed effects model fit to the data however do show a statistically significant effect of region on the absolute measurement of shimmer ($\eta^2 = 0.023$, $p < 0.01$).

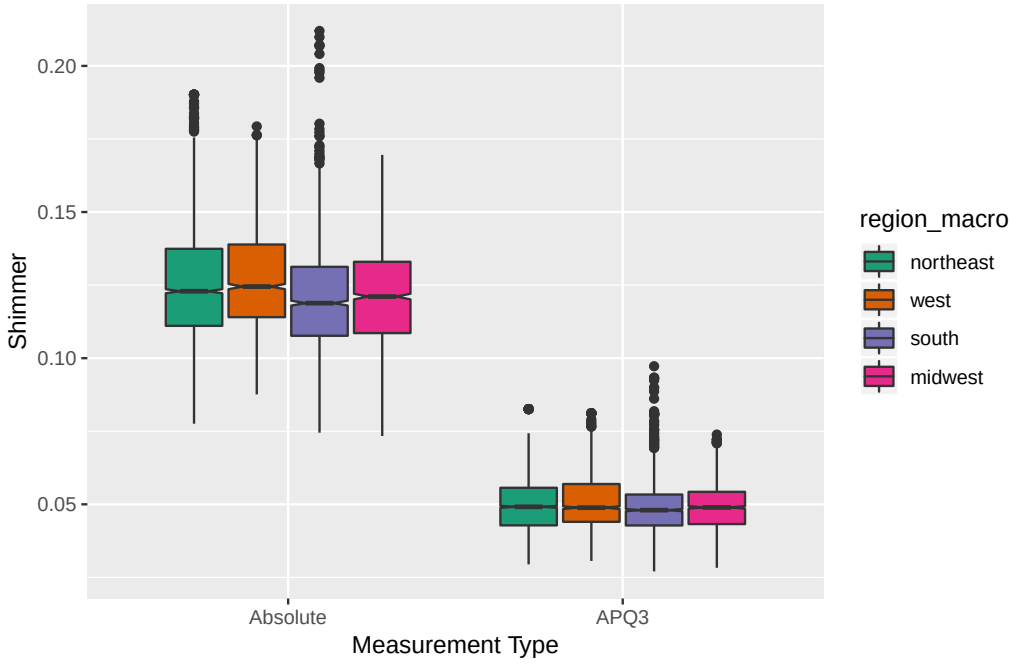

Figure 4.55: Region differences in shimmer at the segment level

4.4.2.4 Pitch

While the differences in min pitch and max pitch appear negligible, there does appear to be a substantial difference between the region groups in terms of mean pitch. A linear mixed effects model fit to the data confirms a weak yet statistically significant effect of region with respect to mean pitch ($\eta^2 = 0.018$, $p < 0.05$).

As with jitter, the only other sociodemographic trait examined here to exhibit a significant difference in mean pitch is gender, and it appears that gender may well be the underlying cause of the regional differences seen here. Recall from table 4.4 that the northeast and west groups were roughly evenly split between male and female speakers, while the south and midwest groups were 62% and 71% female, respectively. Also recall from section 4.1.2.4 that females exhibited significantly higher values for mean pitch than males. That the two most heavily female region groups would exhibit the highest values of mean pitch is therefore somewhat unsurprising.

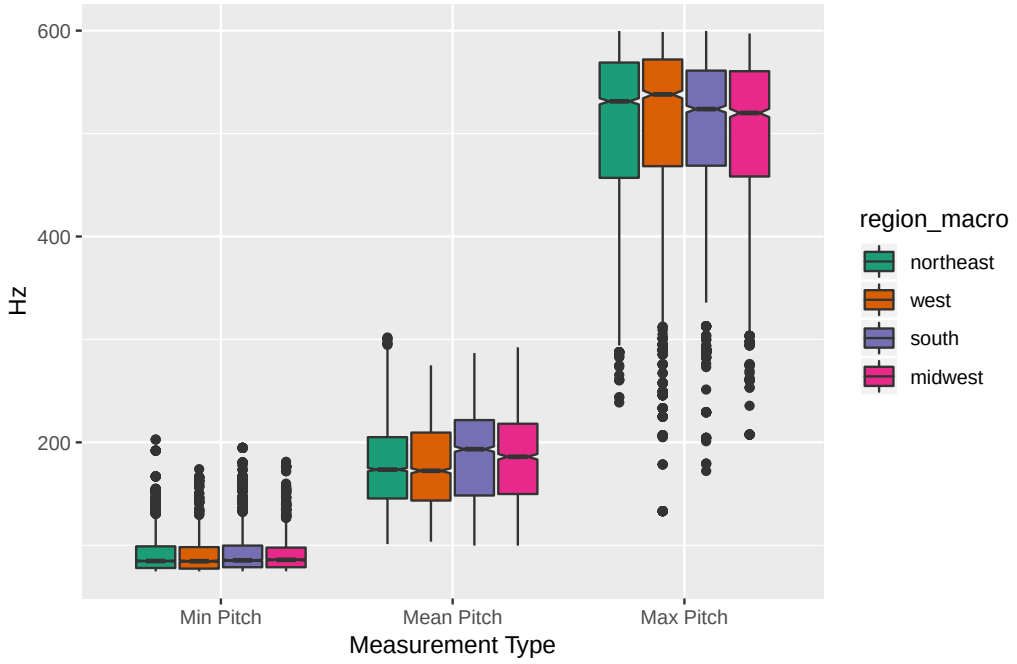

Figure 4.56: Region differences in pitch at the segment level

4.4.3 Phonetic variables

4.4.3.1 Vowel space

Figure 4.57 appears to show little if any effect of region on vowel space area. Likewise, vowel space dispersion does not appear from figure 4.58 to exhibit any meaningful differences with respect to region. Linear mixed effects models fit to the data show the effect of region on these variables to be quite weak and non-significant.

Differences in vowel dynamicity shown in figure 4.59 also appear minimal, however a linear mixed effects model fit to the data does find a significant effect of region on dynamicity ($\eta^2 = 0.018$, $p < 0.05$). This effect is expected, as many dialects in the southern region of the united states have been found to exhibit high degrees of monophthongization of vowel segments which in most other regions are diphthongal (particularly AY; Labov et al., 2006), and conversely several dialects of the northeast

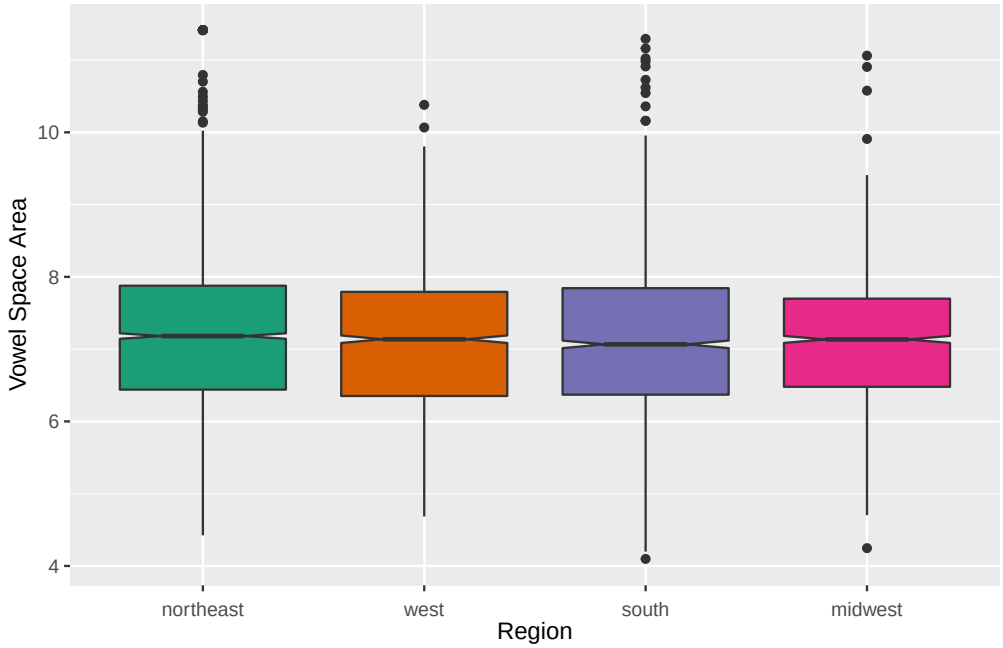

Figure 4.57: Region differences in vowel space area

are characterized by a high degree of movement in what in most other regions are monophthongal vowel segments (particularly AO and AE; Labov et al., 2006).

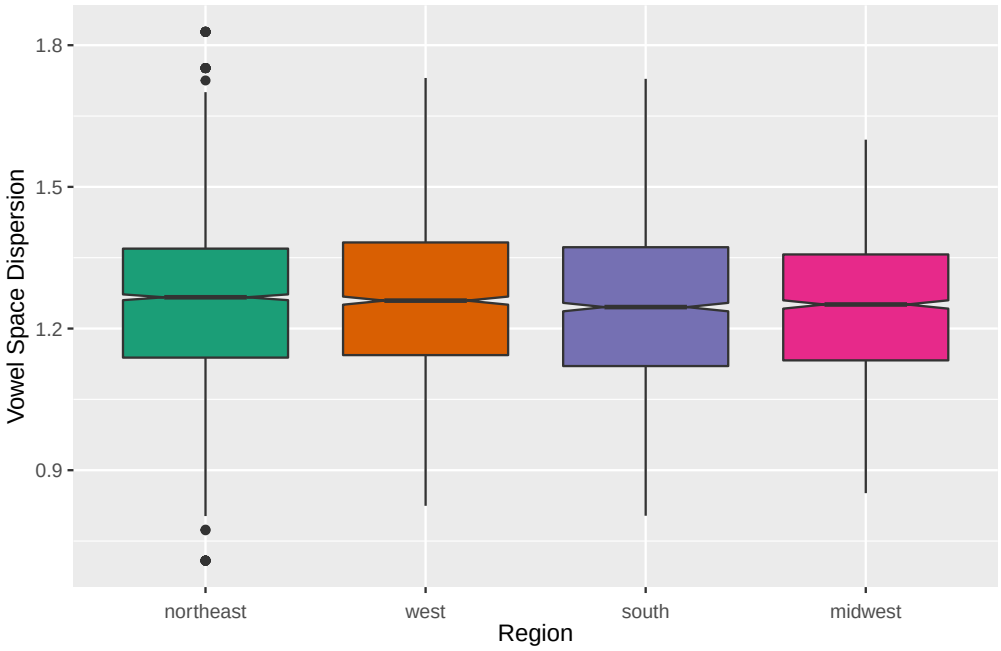

Figure 4.58: Region differences in vowel space dispersion

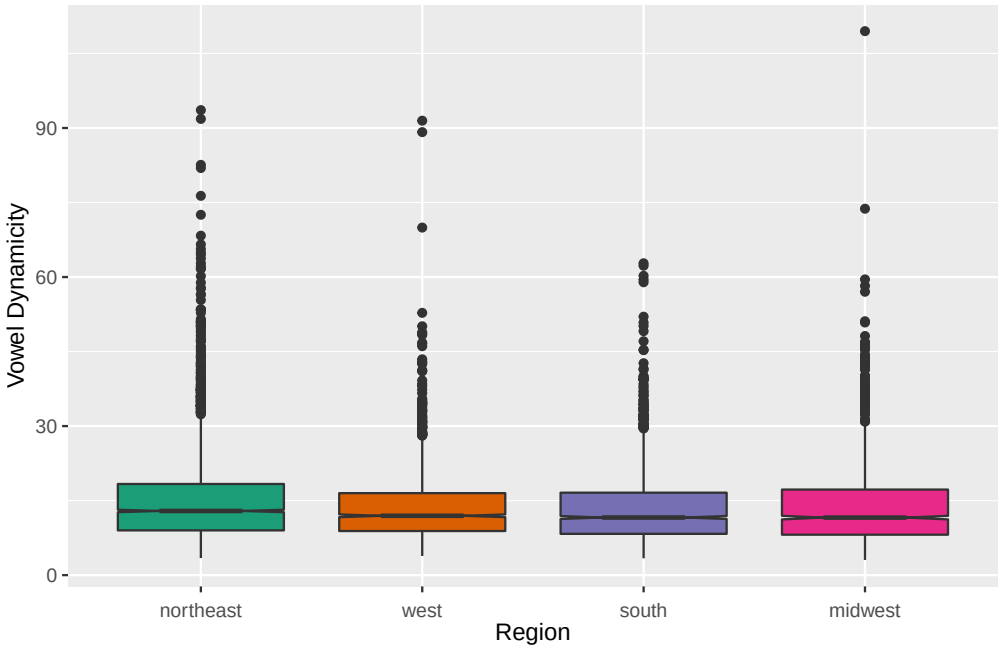

Figure 4.59: Region differences in vowel dynamicity

4.4.3.2 Vowel positions

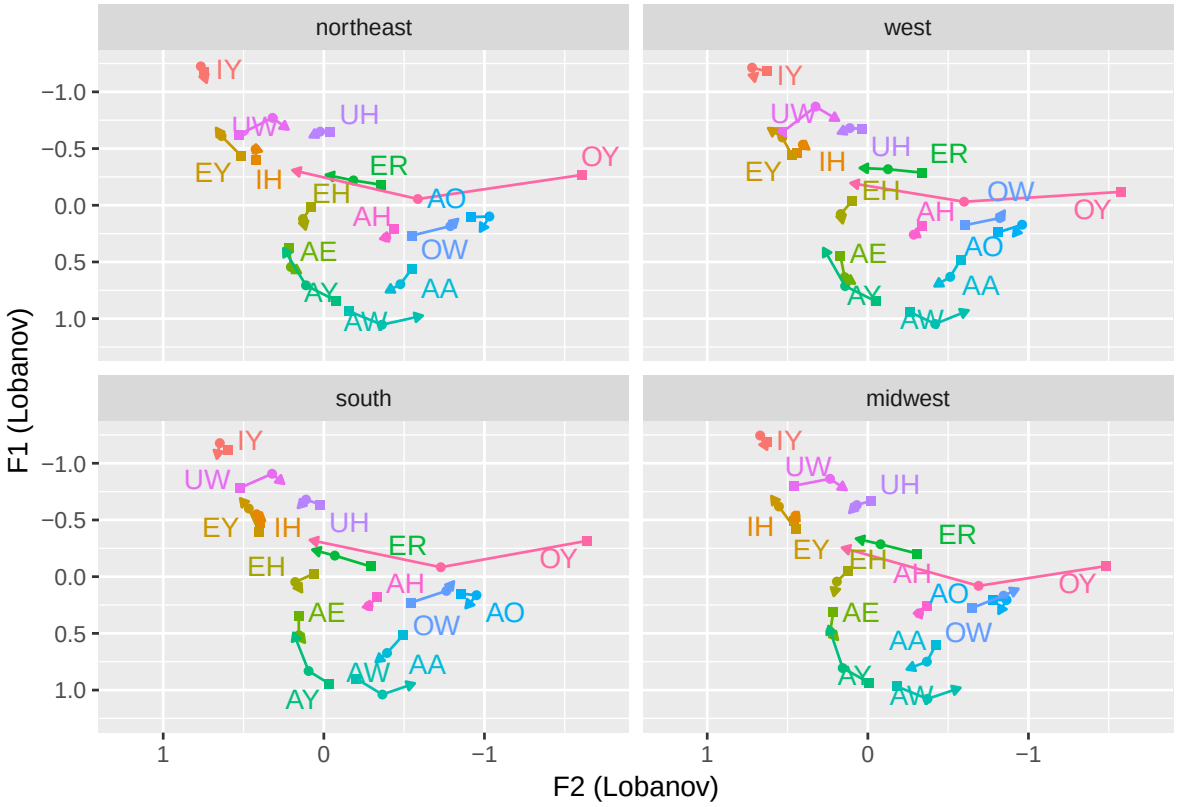

Figure 4.60: Vowel positions and trajectories by region

Figure 4.60 presents the relative positions of the 14 vowels analyzed for each of the four regions. As expected, many of the vowels presented in figure 4.60 appear to exhibit meaningful differences in relative position with respect to region. A few of the more noticeable differences are noted below. Refer to appendix C for a full list of those vowel/point combinations on which region was shown to have a significant effect.

- **OW**: Though similar in dynamicity, the onset of the OW vowel for midwestern speakers appears relatively backed.
- **AH**: Onset and offglide for AH appear slightly backed for northeastern speakers.
- **IY**: Onset of IY appears somewhat fronted for northeastern speakers as compared to the other three groups.
- **AE**: Onset of AE appears slightly lower for western speakers and slightly higher for midwestern speakers than the other two groups.
- **AA**: Onset of AA appears slightly fronted for the midwestern group as compared to the other three groups.

4.4.4 Lexical features

4.4.4.1 Quotatives

Figure 4.61 appears to show salient differences in regional usage for the "say", "be like", and "go" quotatives.

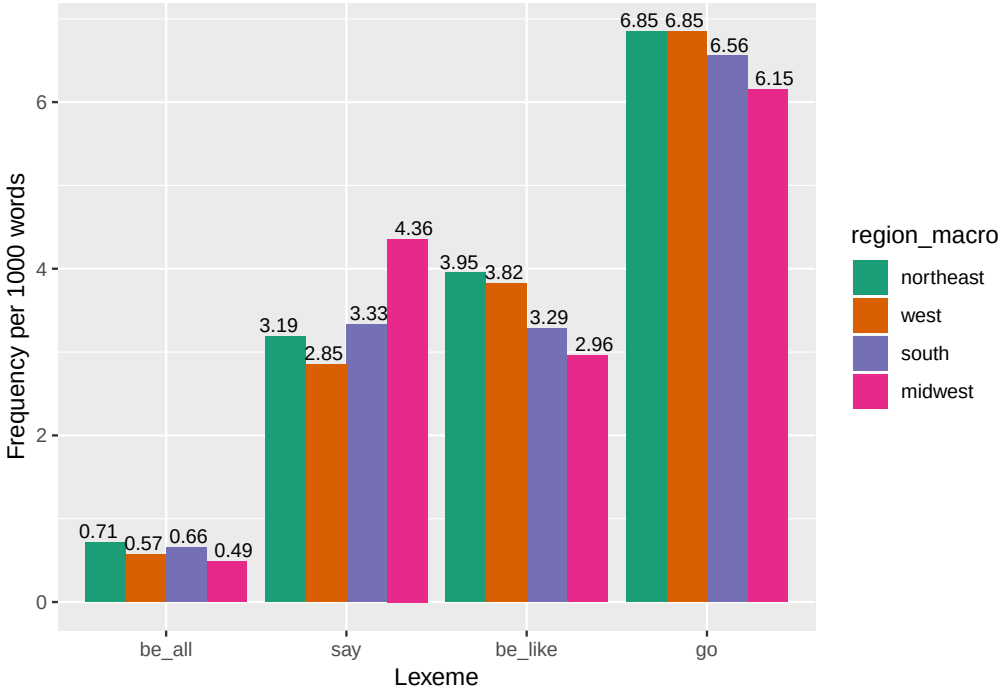

Figure 4.61: Quotative usage by region

Recall that the most significant factor impacting quotative usage found so far has been age, as detailed in section 4.3.4.1. It is useful therefore to examine the age-breakdown within each regional group prior to interpreting the differences found in figure 4.61. Table 4.5 presents the percentage of each regional category made up by each age group.

The west regional group skews the youngest, followed by the northeast, the south, and the midwest. In light of this, it appears that the differences in regional quotative usage may in large part be driven by differential age make-up, with the youngest groups

Table 4.5: Age group percentages by region category.

age_cat	northeast	west	south	midwest
16-25	0.24	0.39	0.15	0.09
26-35	0.30	0.28	0.39	0.28
36-45	0.24	0.17	0.25	0.28
46-55	0.12	0.09	0.16	0.25
56+	0.10	0.07	0.05	0.11

preferring the newer quotatives "go" and "be like," and the older regions preferring the more traditional quotative say. This is consistent with the patterns found in section 4.3.4.1.

4.4.4.2 Modals

Again, the differences in "have to" usage appear to be motivated by age, as this modal construction was shown in section 4.3.4 to be preferred by middle aged speakers, and appears here to be used more by the two more middle-aged regional groups than either the youngest (west) or the oldest (midwest) groups. However, though there appears from figure 4.62 to be several modals for which regional variation may have a meaningful impact, linear mixed effects fit to the data showed no significant effect of region on any of the modal constructions examined here.

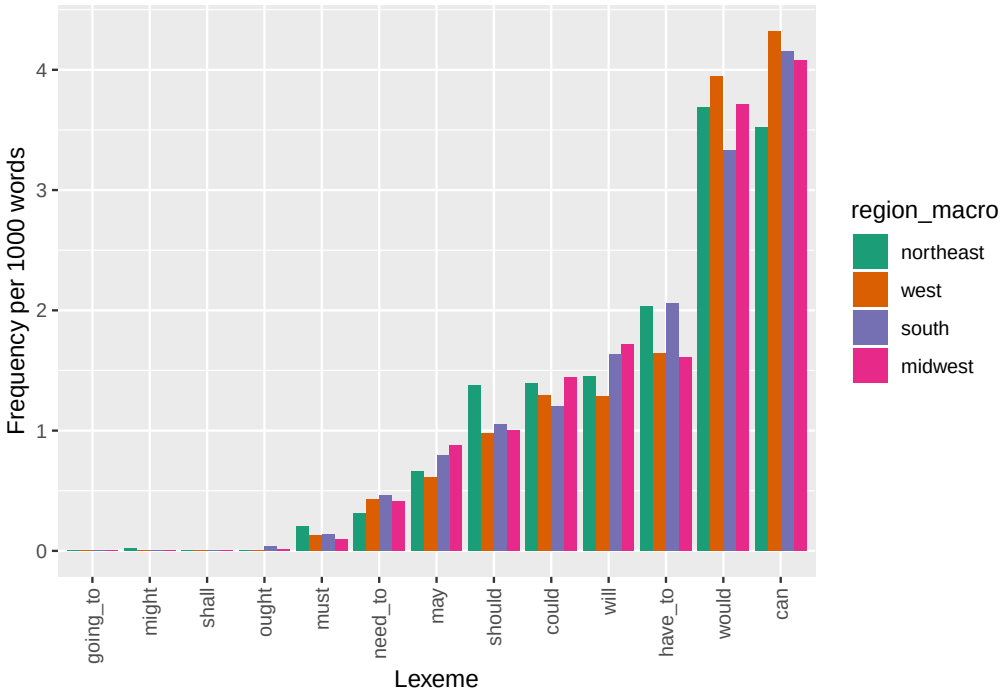

Figure 4.62: Modal usage by region

4.4.4.3 Intensifiers

Figure 4.63 compares rates of intensifier usage among the four regional categories. The only intensifier to show a significant effect of region in linear mixed effects models fit to the data was "really" ($\eta^2 = 0.021$, $p < 0.01$). The intensifier "really" was also shown in section 4.3.4 to be heavily influenced by speaker age, and the regional distribution follows the same pattern– younger regions lead the usage of the "really" intensifier.

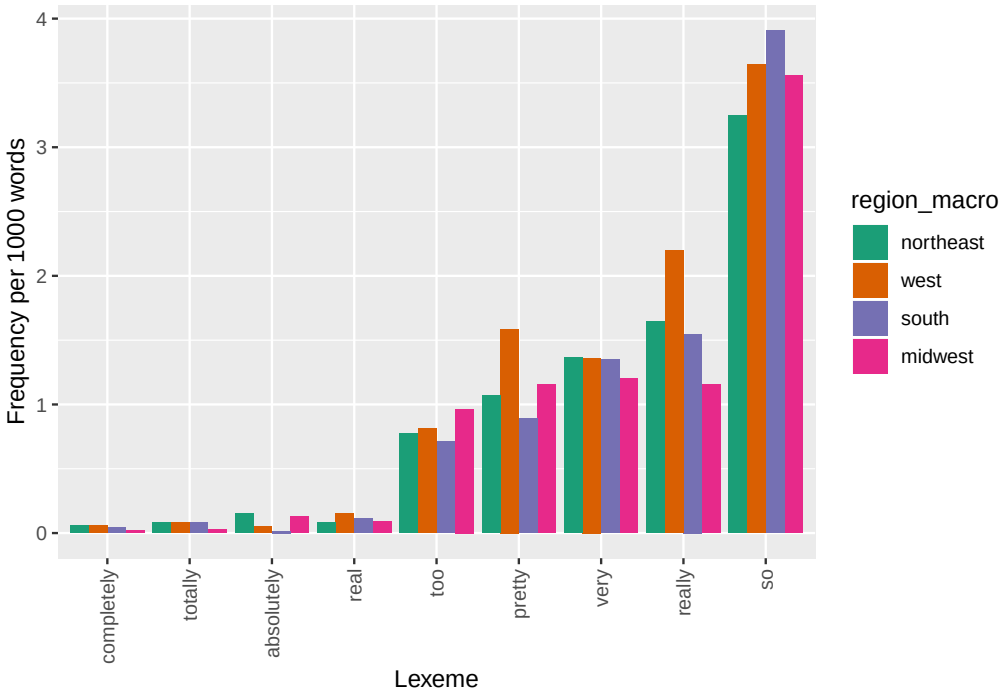

Figure 4.63: Intensifier usage by region

4.4.4.4 Discourse markers

As with the previously discussed sociodemographic traits, regional variation appears from figure 4.64 to be most prevalent in the discourse markers "yeah" and "like," both of which show significant effects of region in linear mixed effects models fit to the data. Again, usage of these discourse markers appears likely influenced by within-group age make-up, with the youngest groups far outstripping the oldest groups in frequency of use.

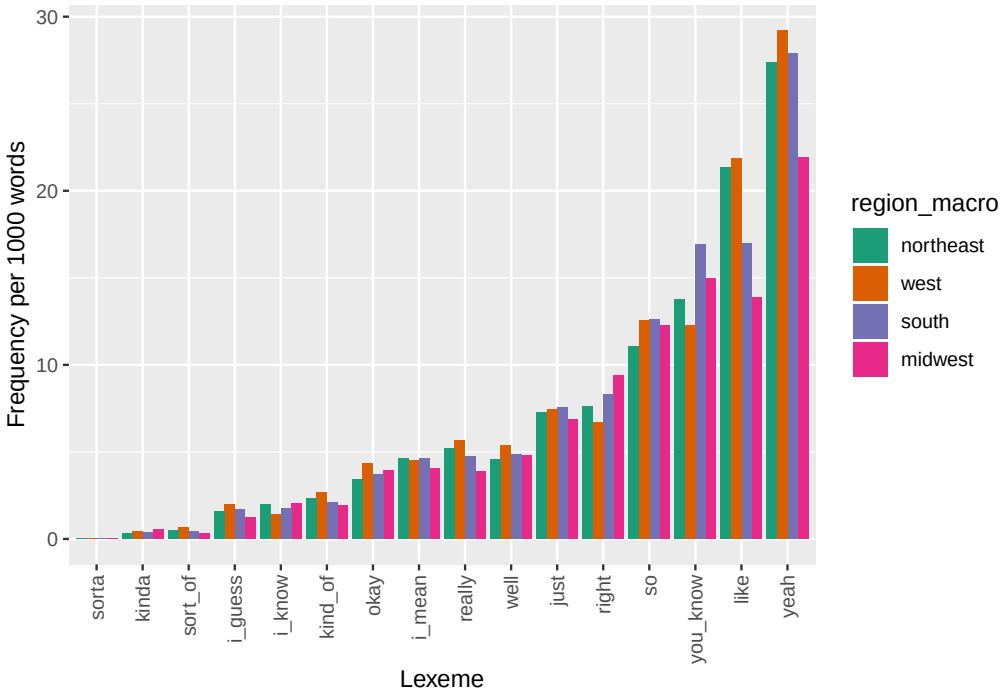

Figure 4.64: Discourse marker usage by region

4.4.4.5 Pronoun proportions

As with discourse markers, regional variation in pronoun usage follows the pattern observed for age groups in section 4.3.4, with older regional groups using first person pronouns less frequently and third person pronouns more frequently than younger regional groups. While the difference among age groups did not reach significance in first pronoun usage, a linear mixed effects model fit to the data does show a statistically significant effect of region on third person pronoun usage ($\eta^2 = 0.016$, $p < 0.05$).

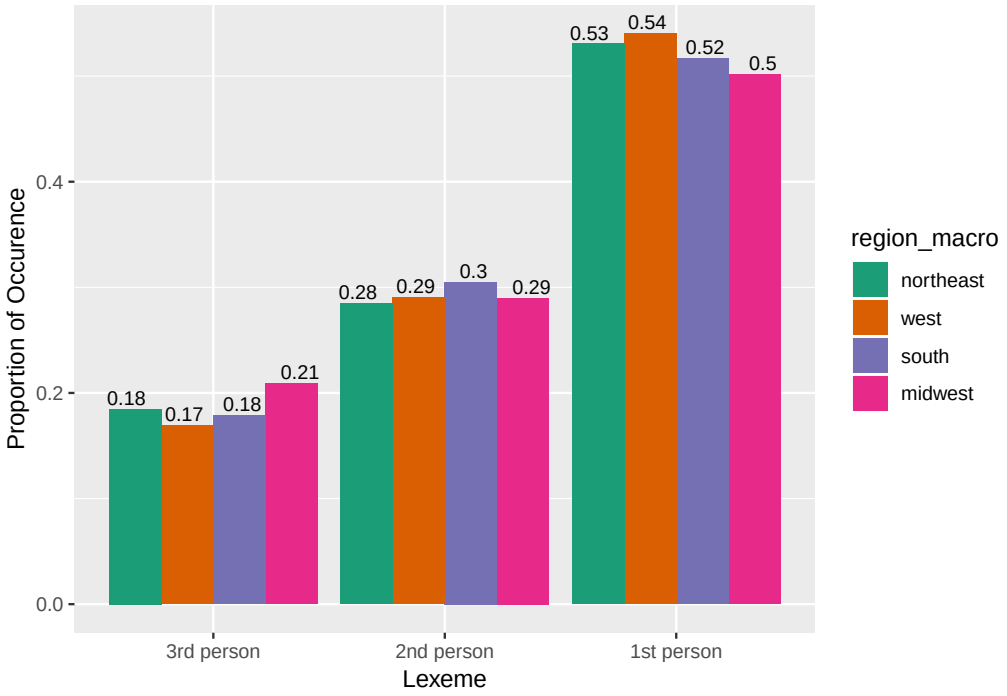

Figure 4.65: First, second, and third person pronoun usage proportions

4.4.4.6 Politeness and taboo frequency

Though there appear to be meaningful regional differences in taboo and politeness frequency in figure 4.66, none of these reached significance in linear mixed effects models fit to the data when taking including a random intercept of speaker ID.

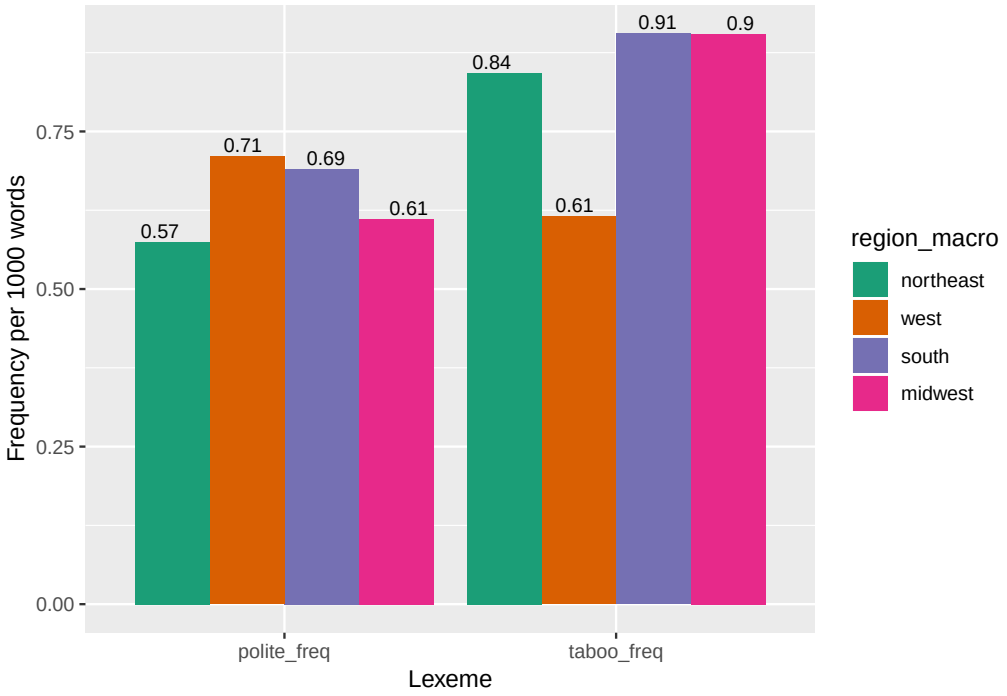

Figure 4.66: Usage of taboo and politeness terms by region

4.4.4.7 Sentiment

The differences in polarity and subjectivity presented in figure 4.67 appear negligible.

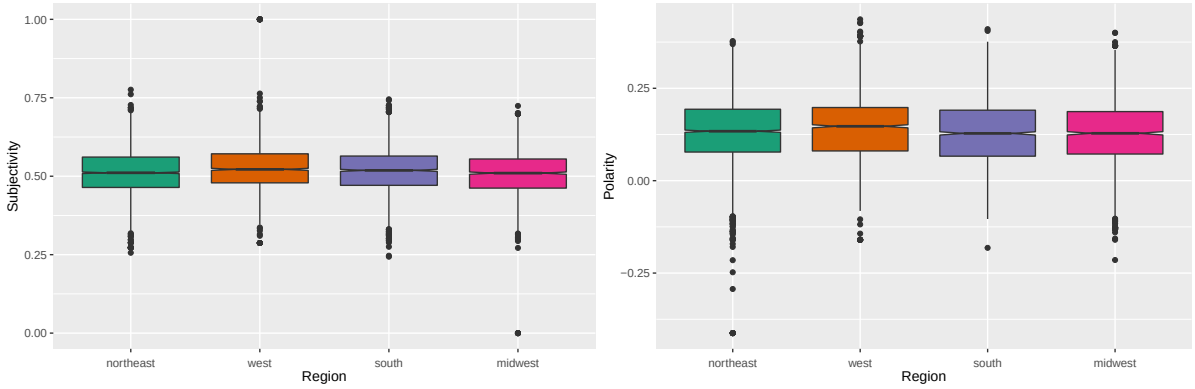

Figure 4.67: Segment subjectivity (left) and polarity (right) by region

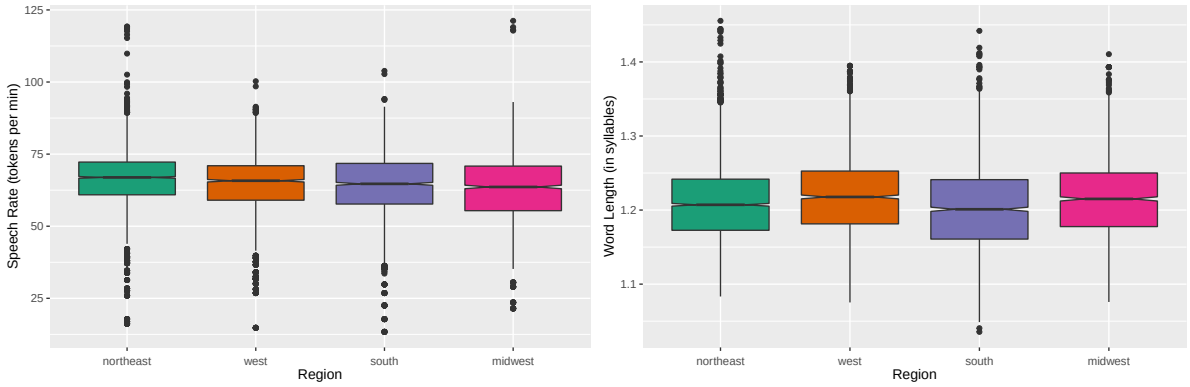

Figure 4.68: Speech rate (left) and word length (right) by region

4.4.4.8 Speech rate and word length

Speakers from the northeast appear to exhibit the fastest speech rate on average, and speakers from the midwest appear to exhibit the slowest. Speakers from the west and midwest exhibit the highest average word lengths, while speakers from the northeast and the south exhibit the lowest average word lengths. Linear mixed effects models confirm a weak but significant effect of region on both word length ($\eta^2 = 0.019$, $p < 0.05$) and speech rate ($\eta^2 = 0.014$, $p < 0.05$).

4.4.4.9 Top informative ngrams

As detailed in chapter 3, the top 2000 ngrams (unigrams and bigrams combined) according to information gain for each sociodemographic trait of focus were extracted to serve as predictor features for the models presented in chapters 5 and 6. Table 4.6 presents the top 20 ngrams according to information gain for speaker region. The full 2000 are not shown here for reasons of space. Unsurprisingly, these tend to almost exclusively contain geographical references, and should be quite helpful to the models presented in chapters 5 and 6 for distinguishing speaker region.

Table 4.6: Top 20 informative ngrams for region

attributes	infogain
berkeley	0.0402464
philadelphia	0.0302372
california	0.0261122
philly	0.0218877
chicago	0.0210857
huge	0.0210818
texas	0.0202634
new_york	0.0200970
york	0.0200970
yeah_i	0.0197340
bay	0.0180502
activity	0.0177537
be_right	0.0176040
southern_california	0.0164826
indiana	0.0162447
texas_i	0.0161463
northern_california	0.0149726
jersey	0.0147293
europe	0.0143708
bay_area	0.0141599

4.5 Education Level

4.5.1 Overall education level distribution

Figure 4.69 shows a histogram overlaid with a density plot of the education years reported by speakers in this data-set.

By far, the most common self-reported number of education years is 16, the typical number of education years experienced by a student in the American education system who has completed elementary through high-school (12 years) and received an undergraduate degree of some sort (4 years). The majority of speakers fall between 12

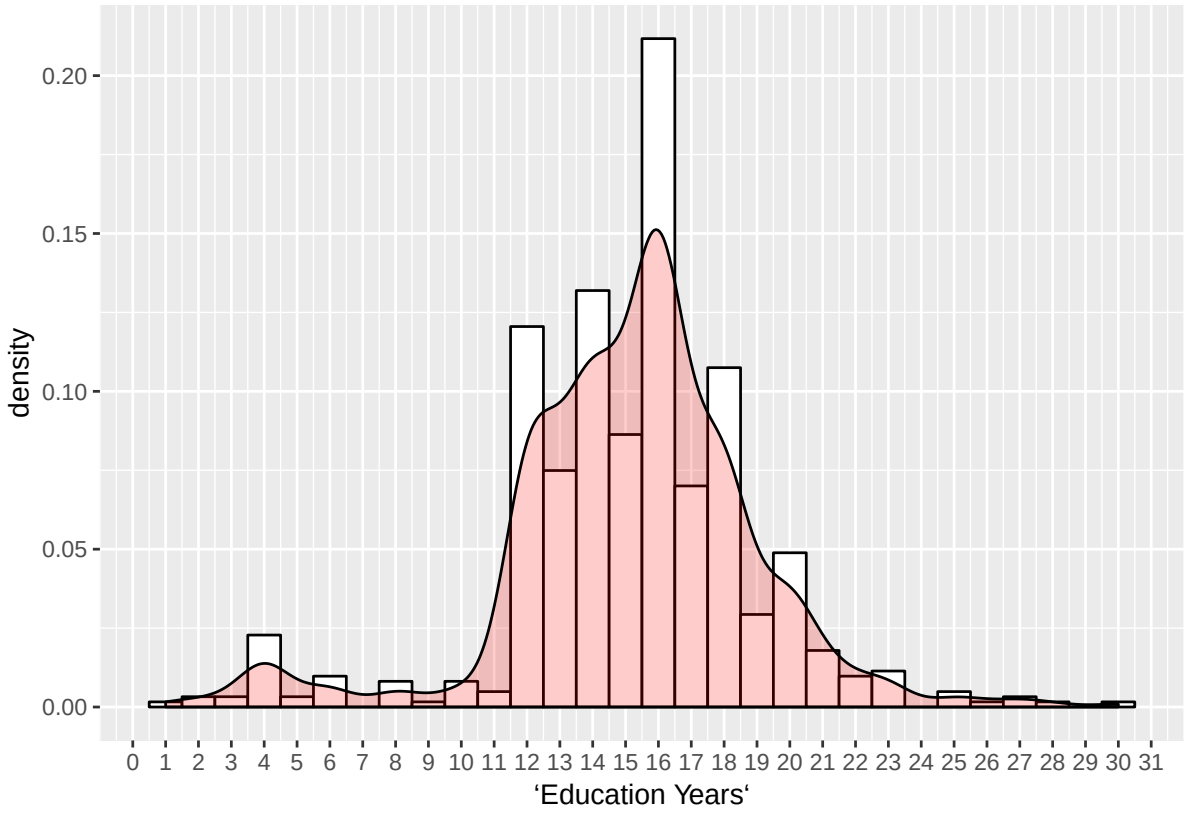

Figure 4.69: Overall distribution of education level within the dataset

years of education (the typical amount for having completed high-school) and 18 years of education (The typical amount for having completed an undergraduate degree plus a Master's or similar post-baccalaureate degree). There are relatively small tails in the distribution of speakers who have completed 11 years or fewer, and 18 years or more. Those who have undergone 11 or fewer years of education are likely those speakers who did not complete secondary schooling. Those with 18 or more years of education are likely those speakers who underwent some level of graduate education after completing an undergraduate degree (and in the case of those with more than 25 years of education, likely multiple post-baccalaureate degrees or particularly intensive medical fellowships and residencies).

As mentioned in chapter 3, in many respects, it makes more sense to group speakers into education categories based on the major delineations present in the American

education system than to treat education as a continuous variable. The plots in figure 4.70 display the breakdown of speakers in this data-set that fall into each of the three education categories laid out in chapter 3 at both the segment and the speaker grouping levels. 51 speakers did not self-report their number of education years and thus are placed in the "NA" category and not included in the following analyses.

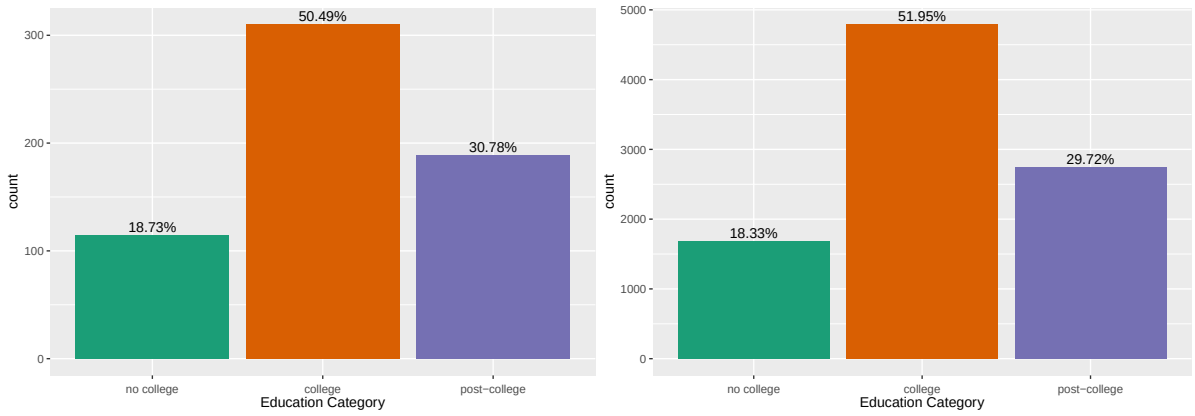

Figure 4.70: Education category breakdown at the speaker (left) and segment (right) levels

There appears to be little meaningful difference in the category distributions at the speaker vs. the segment level, so all following analysis will focus solely on the segment level.

4.5.2 Acoustic variables

4.5.2.1 HNR

Figure 4.71 suggests a slight tendency for speakers in the post-college group to exhibit lower HNR values on average than the other three groups. However, a linear mixed effects model fit to the data does not show a significant effect of education group on HNR values.

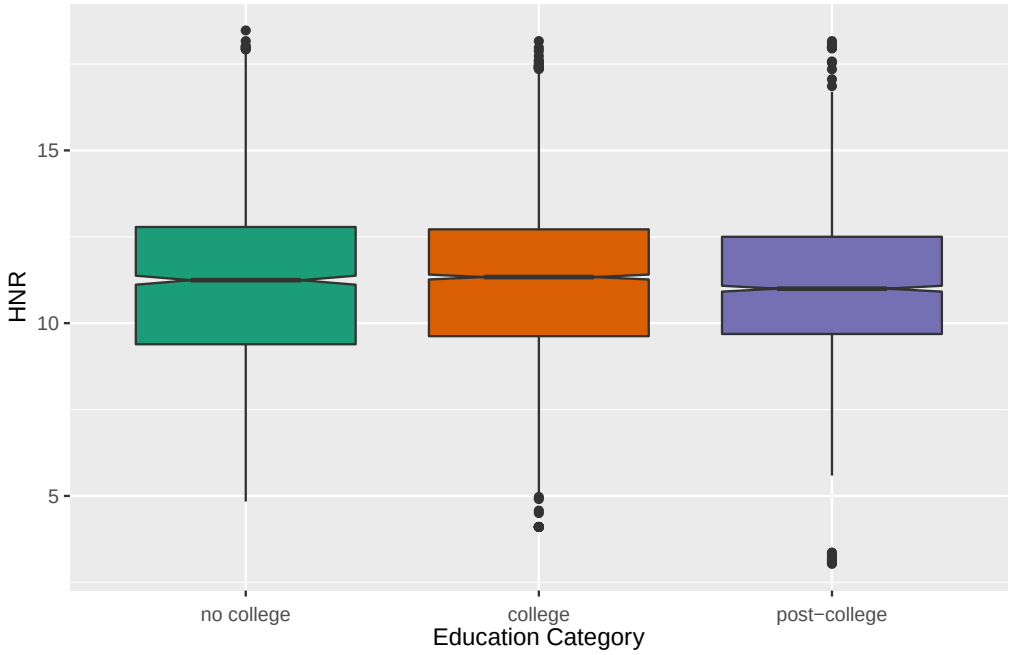

Figure 4.71: Education category differences in HNR

4.5.2.2 Jitter

Figure 4.72 suggests that the college education group tends to have slightly lower jitter values than the other education groups. As with HNR however, a linear mixed effects model fit to the data found no significant effect of education category on any measure of jitter.

4.5.2.3 Shimmer

Figure 4.73 shows little meaningful difference between the education groups with respect to shimmer, and as with the other acoustic variables examined so far, a linear mixed effects model fit to the data found no significant effect of education category on any measure of shimmer.

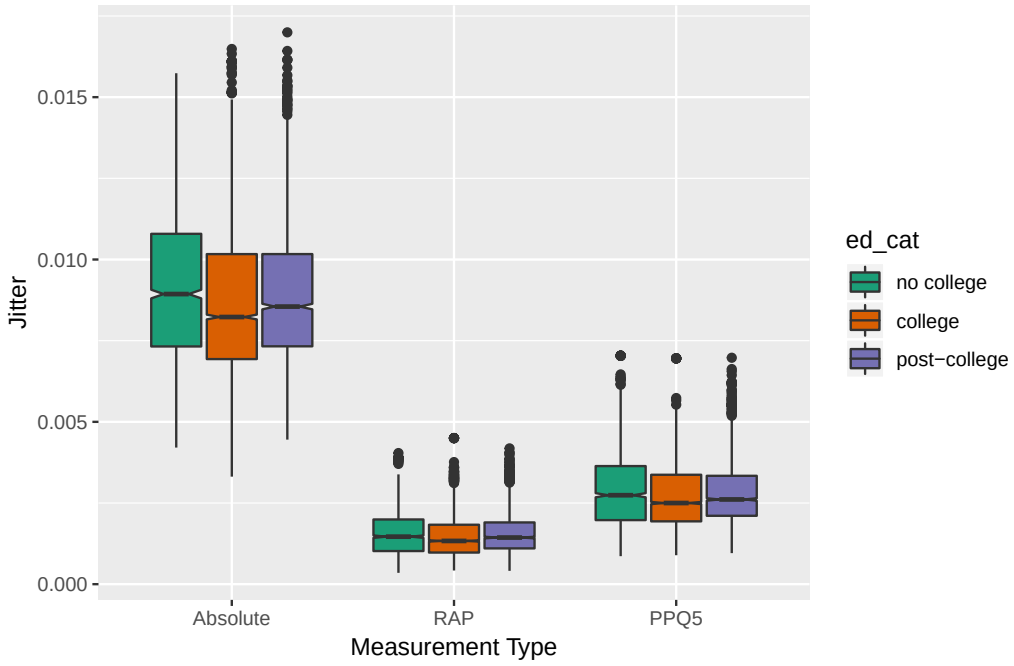

Figure 4.72: Differences in jitter measurements by education category

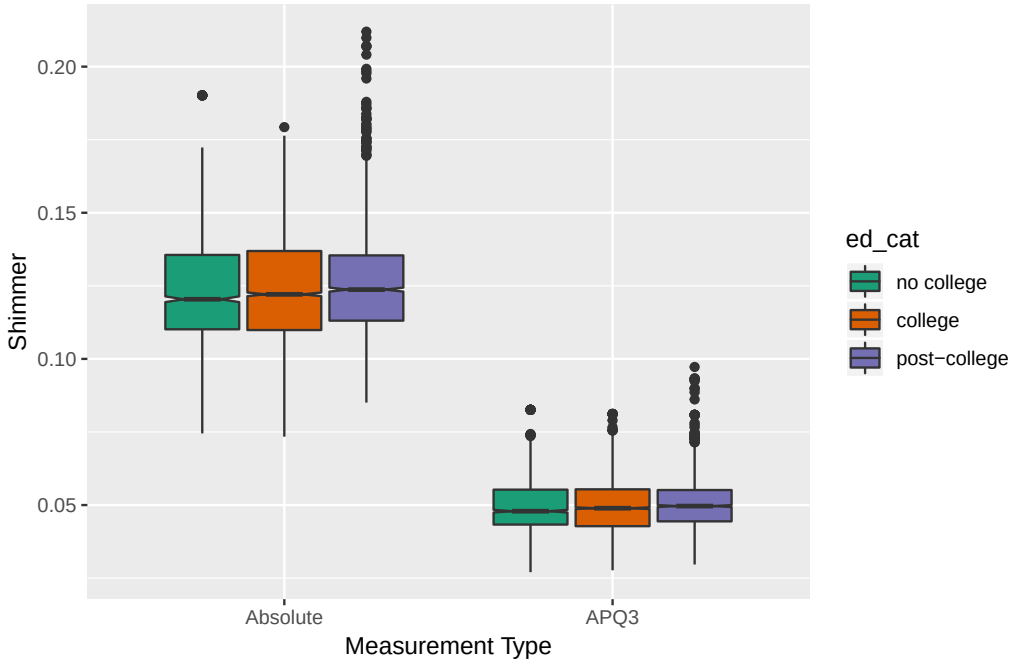

Figure 4.73: Differences in shimmer measurements by education category

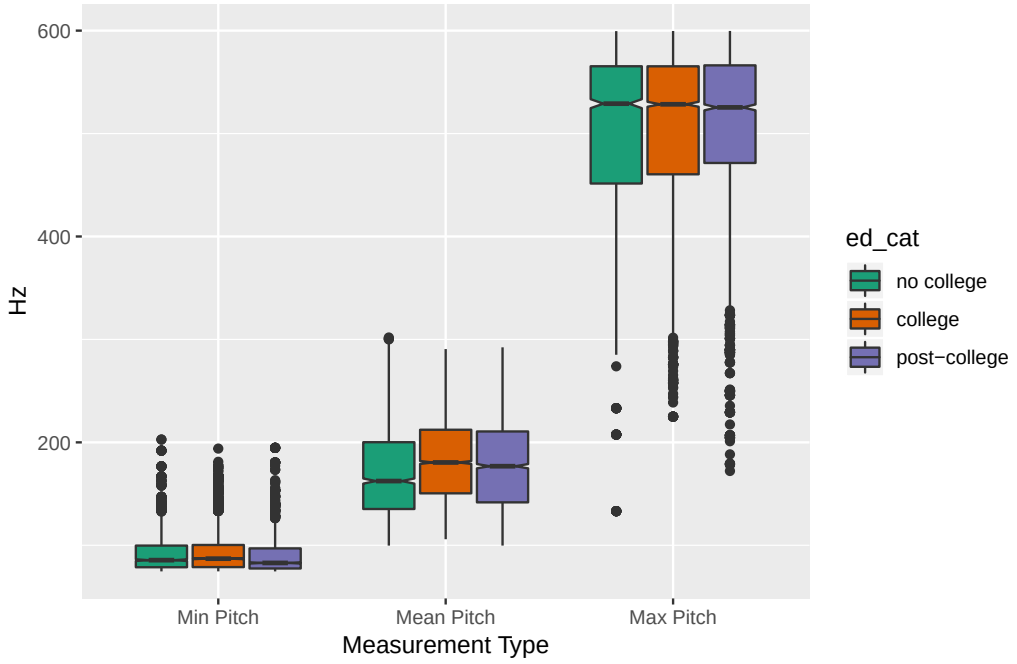

Figure 4.74: Pitch differences between education groups

4.5.2.4 Pitch

Figure 4.74 shows little if any difference in minimum or maximum pitch values, but there does appear to be a relatively large difference in mean pitch between the "no college" group and the other two groups.

That education category would have any sort of effect on pitch values is unexpected, and I know of no literature which suggests that this might be so. It is likely that education group in this instance is acting as a proxy for some other variable that meaningfully impacts speaker pitch. Recall from section 4.1.2.4 that females on average exhibit significantly higher mean pitch values than do males. Table 4.7 shows that the "no college"" group is predominantly male whereas the other two groups are predominantly female. It seems likely therefore that the trend observed in figure 4.74 is driven by the differential sex distribution among the three education categories.

Table 4.7: Sex category percentages by education category

Sex	no college	college	post-college
female	0.42	0.58	0.53
male	0.58	0.42	0.47

However, despite what appears to be a clear difference between the "no college" group and the other two groups, a linear mixed effects model fit to the data found no significant effect of education category on mean pitch.

4.5.3 Phonetic variables

4.5.3.1 Vowel space

Figures 4.75 through 4.77 show effectively no difference in any of the vowel space measures among the different education categories. This is confirmed via linear mixed effects models fit to the data, which found no significant effect of education category for any of these features.

4.5.3.2 Vowel positions

Figure 4.78 displays vowel positions and trajectories for all 14 vowels analyzed for each of the four education categories.

The most salient differences in vowel position and trajectory with respect to education category are discussed in detail below. Refer to appendix C for a list of those vowel/point combinations on which education category was shown to have a significant effect.

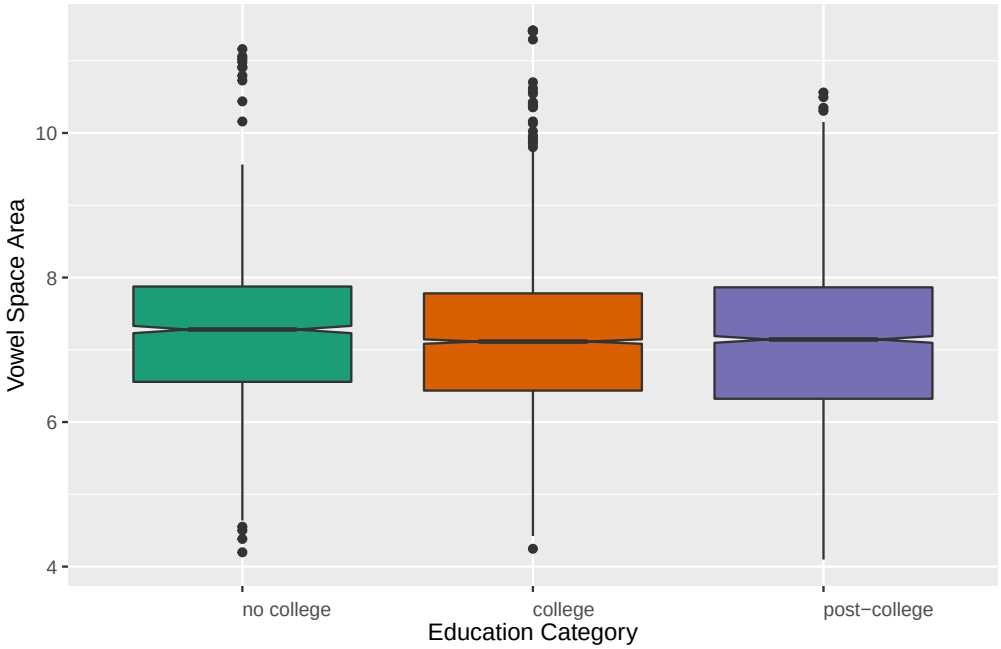

Figure 4.75: Education differences in vowel space area

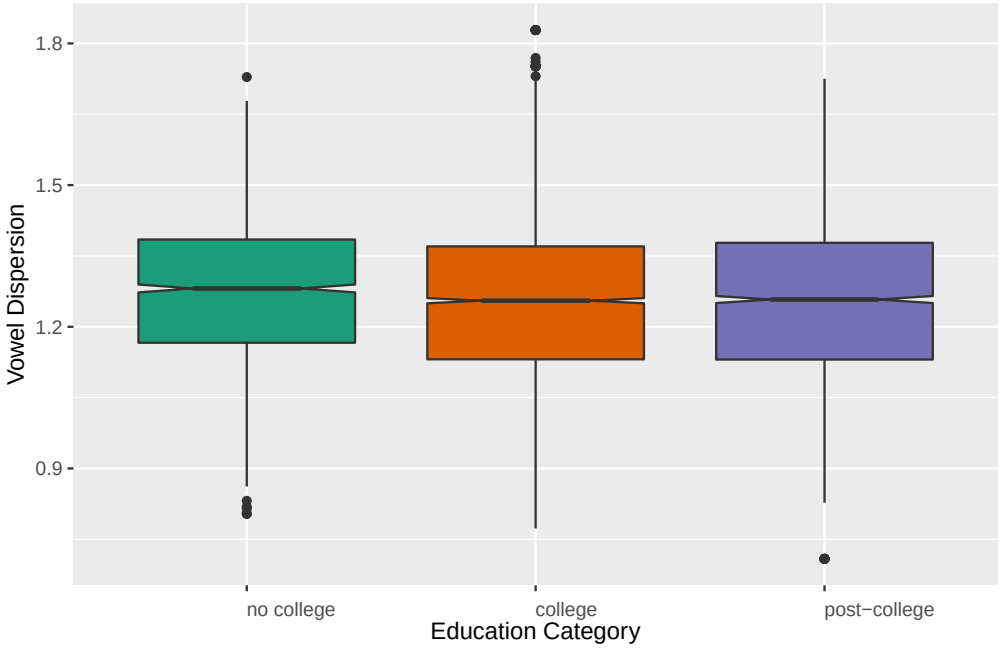

Figure 4.76: Education differences in vowel space dispersion

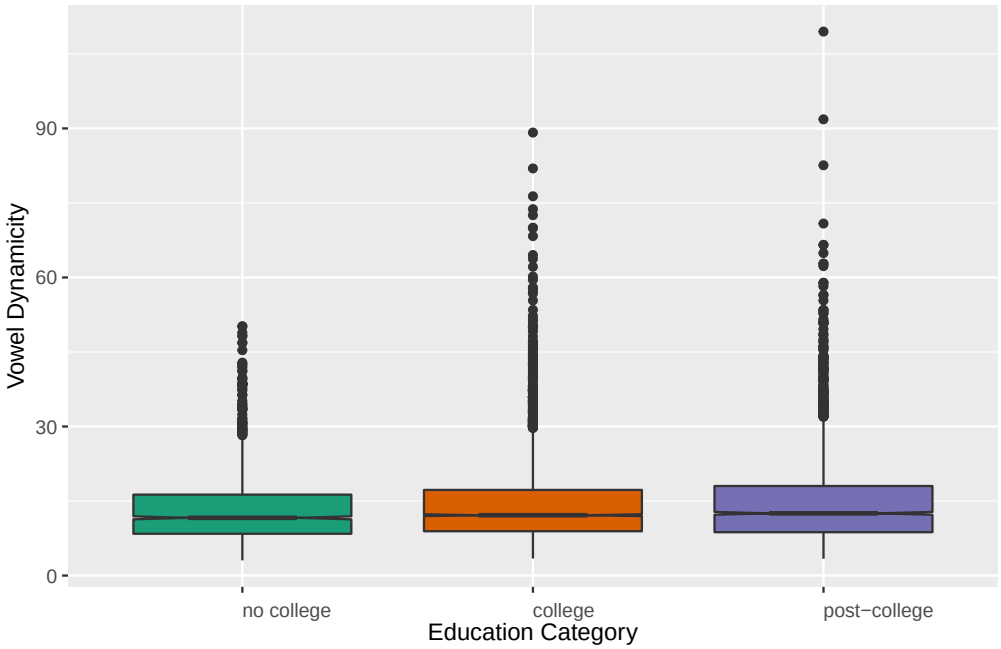

Figure 4.77: Education differences in vowel dynamicity

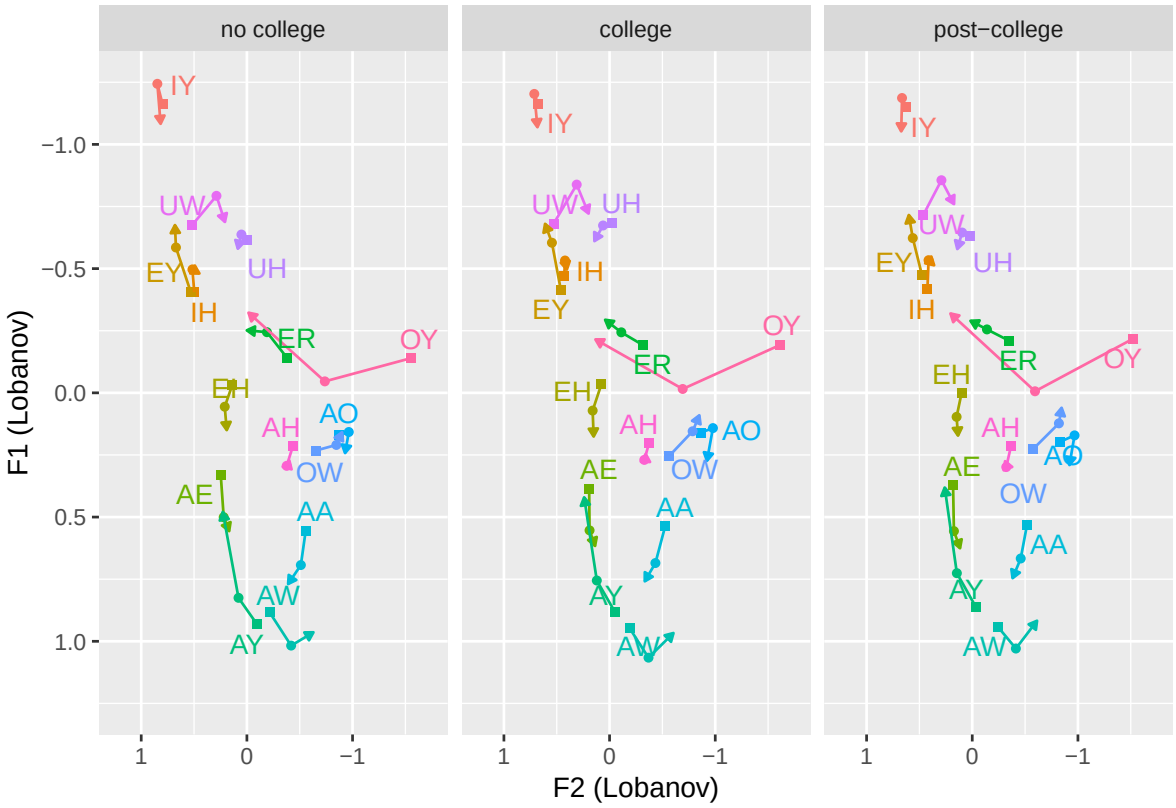

Figure 4.78: Vowel trajectories by education category

- **OW**: The "college" and "post-college" groups appear to realize OW with a more fronted onset than the other group.
- **AE**: The "no college" group appears to realize the entire AE trajectory higher in the vowel space than do the other two groups.
- **UH**: UH for the "no college" group is extremely monophthongal compared to the other three groups.

4.5.4 Lexical features

4.5.4.1 Quotatives

Figure 4.79 shows the average kword frequency of use for the four quotatives considered here.

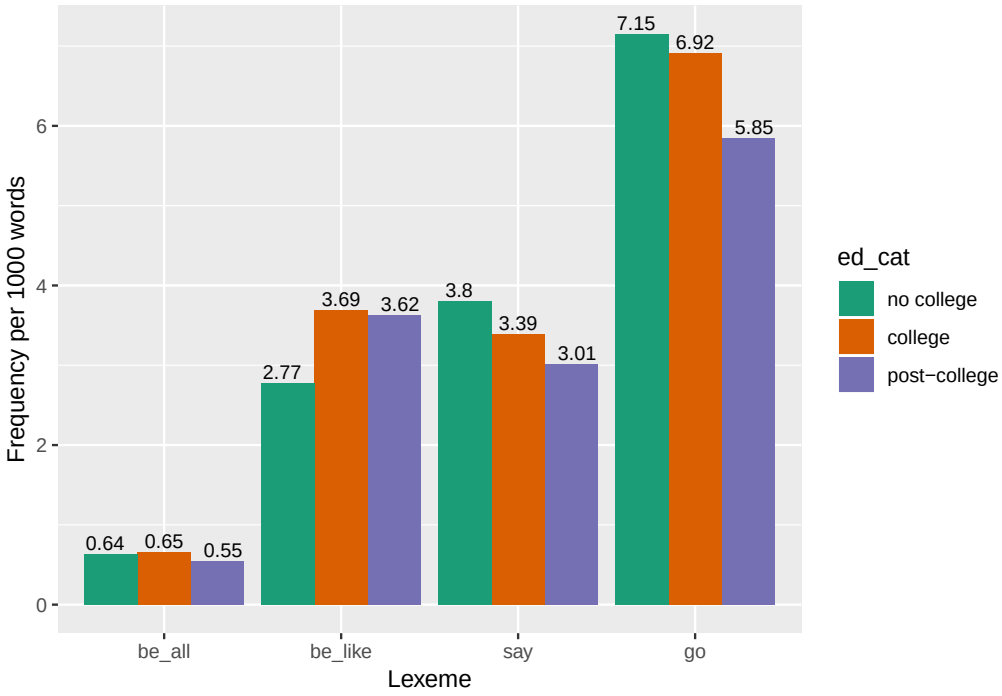

Figure 4.79: Quotative usage by education

The pattern of quotative use among the education categories looks quite similar to the patterns observed for quotative use with respect to age. Table 4.8 provides a comparison of the percentages of each education group that are made up of each age

Table 4.8: Age group percentages by education category

age_cat	no college	college	post-college
16-25	0.15	0.26	0.22
26-35	0.19	0.26	0.44
36-45	0.30	0.23	0.19
46-55	0.26	0.15	0.08
56+	0.11	0.10	0.07

group.

Generally, it appears that education can function somewhat as a proxy for age in this corpus, with more educated speakers on average belonging to younger age groups, and less educated speakers on average belonging to older age groups. In light of this, it is unsurprising that more education corresponds to higher use of "be like" and lower use of "say," given the findings related to quotative usage by age in section 4.3.4.

It should be noted however that a linear mixed effects model fit to the data found a significant effect of education category only for frequency of the quotative "go" ($\eta^2 = 0.014$, $p < 0.05$).

4.5.4.2 Modals

Figure 4.80 shows the distribution of the examined modal constructions with respect to education category.

Interestingly, while the usage patterns for modals with respect to education category does in many cases mirror the usage pattern seen for age, there are a few deviations. "Can" appears to be the modal item that deviates the most from what we would expect given the findings in section 4.3.4 and the distribution of age within the education categories. It appears from figure 4.80 that usage of "can" increases with education

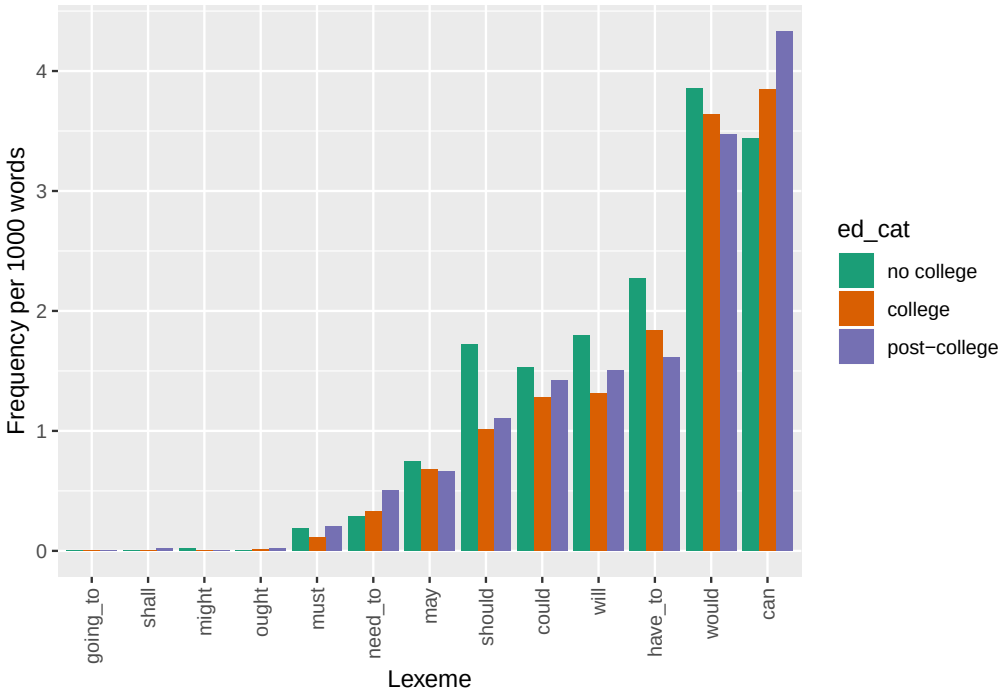

Figure 4.80: Modal usage by education

level, though a corresponding increase in the use of "can" was not observed for younger speakers. Nor does this appear to be related to the distribution of ethnicities within the different age groups, as little variation among ethnicities was found in section 4.2.4 with regard to frequency of "can." There may be a bit of a sex effect here, as females were found in section 4.1.4 to use "can" slightly more than males on average, but the difference between the highest two education groups here and the lowest two is larger than that seen for sex. This may be a legitimate difference related to level of education, though why higher levels of education would lead to different usage patterns for specific modal constructions is unclear.

4.5.4.3 Intensifiers

As with quotatives, the usage pattern for intensifiers among the education groups largely mirrors that for age groups. Speakers in the lower education levels (older speakers) favor

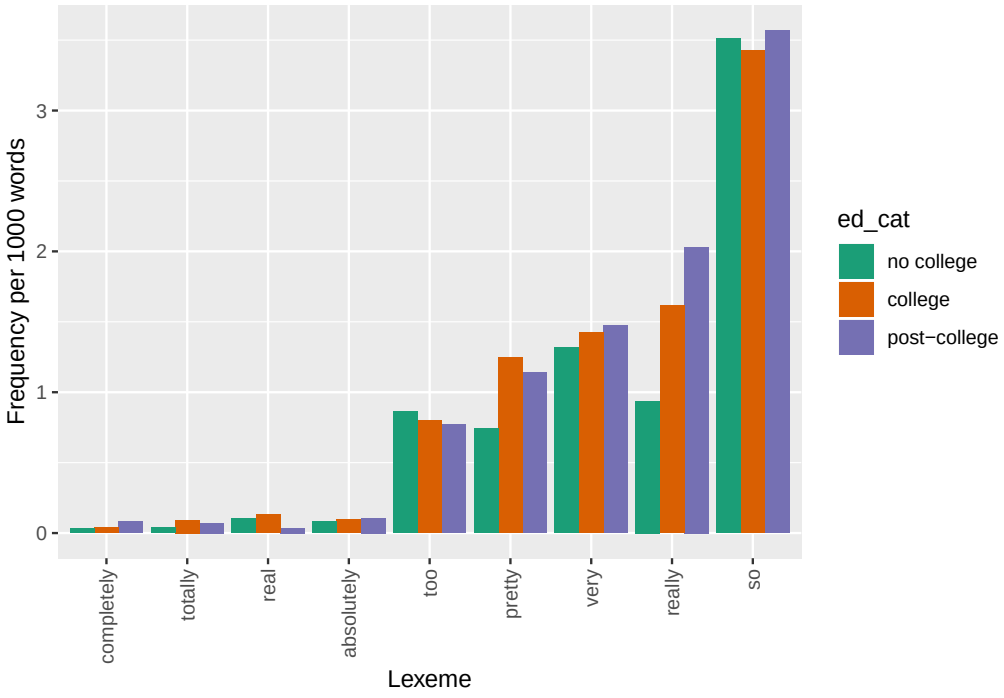

Figure 4.81: Intensifier usage by education

"too" and "real" while speakers in the higher education groups (younger speakers) favor "pretty" and "really." Interestingly, it seems that more educated speakers slightly favor "very" while in section 4.3.4 very was found to be favored heavily by the oldest age group. However, no significant effect of education group on was found in a linear mixed effects model fit to the data for any intensifier except "really" ($\eta^2 = 0.023$, $p < 0.001$) and "pretty" ($\eta^2 = 0.013$, $p < 0.05$).

4.5.4.4 Discourse markers

As with quotatives and modals, the pattern of usage for discourse markers among education groups largely follows what we would expect given their age makeup. Those discourse markers favored by younger speakers are also favored by speakers with high levels of education, and those favored by older speakers are also favored by those with low levels of education. The only obvious counter examples of this are the high usage of

"yeah" by speakers in the "no college" category, and the relatively even usage of "just" and "so" by all education groups.

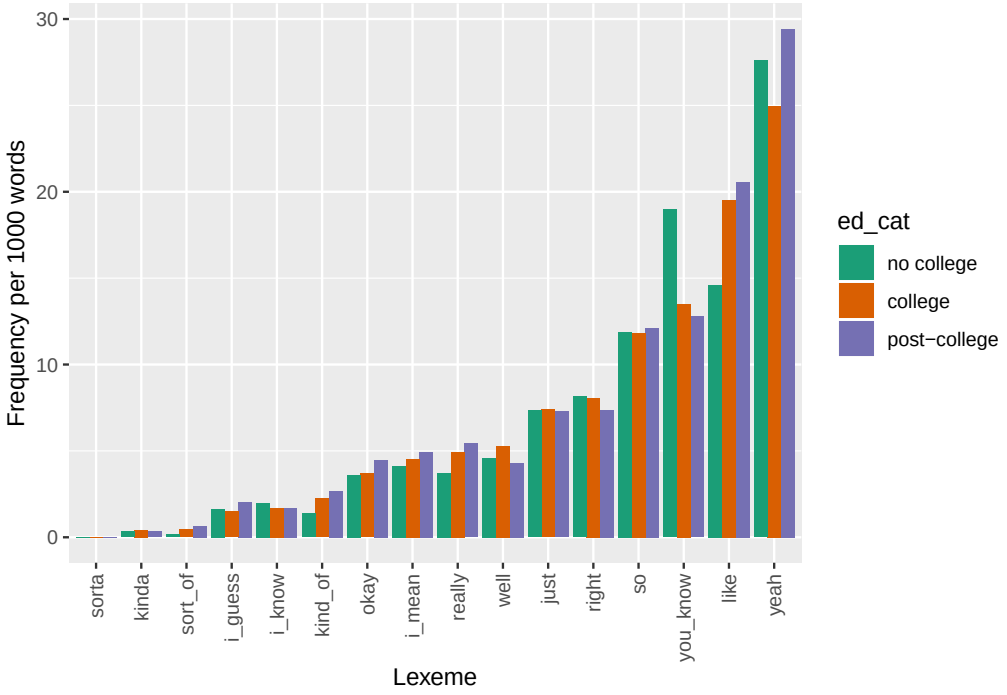

Figure 4.82: Discourse marker usage by education

Linear mixed effects models fit to the data show significant effects of education category for the discourse markers "really" ($\eta^2 = 0.022$, $p < 0.01$), "like" ($\eta^2 = 0.012$, $p < 0.05$), "kind of" ($\eta^2 = 0.021$, $p < 0.01$), "sort of" ($\eta^2 = 0.012$, $p < 0.05$), and "you know" ($\eta^2 = 0.014$, $p < 0.05$).

4.5.4.5 Pronoun proportions

Figure 4.83 shows the proportion of pronoun occurrence observed with respect to education category.

Again mirroring expectations based on mean group age, the education groups with the youngest mean age also exhibit the highest rates of first person pronoun usage and the education groups with the oldest mean age exhibit the highest rates of third

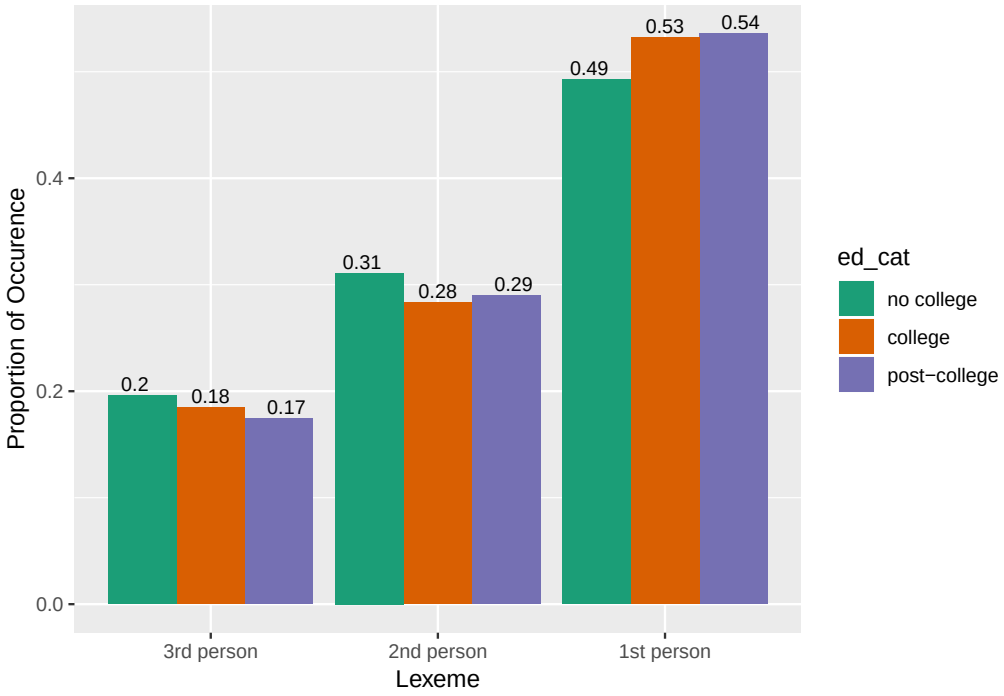

Figure 4.83: First, second, and third person pronoun usage proportions

person pronoun usage. Linear mixed effects models fit to the data found no significant effects of education category on pronoun usage, however.

4.5.4.6 Taboo and politeness terminology

Politeness terminology patterns for education category again seem to mirror the patterns found for age. Politeness terminology was more frequent in the youngest and oldest speaker groups, and appears here to be highest in the youngest and oldest education groups. Interestingly, the trend observed above for education groups to behave as one would expect given their respective age makeups does not hold for production rates of taboo terms. Whereas taboo word usage is highest for the youngest and the oldest speaker groups and lowest for those middle-age groups, taboo usage across education levels seems to decline with higher education.

Despite the apparent trends observed here, no significant differences were found

during linear mixed effects model testing between education groups for either of these predictors.

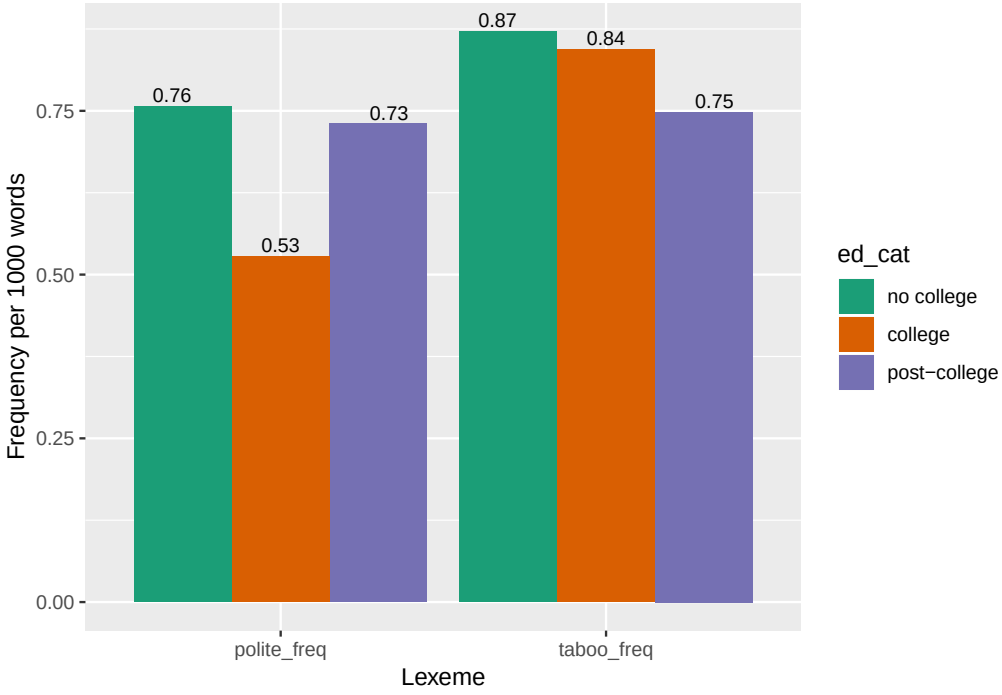

Figure 4.84: Usage of taboo and politeness terms by education

4.5.4.7 Sentiment

There appears from figure 4.85 to be no meaningful difference between education groups in terms of polarity or subjectivity. This was confirmed via linear mixed effects models fit to the data.

4.5.4.8 Speech rate and word length

Figure 4.86 compares word length and speech rate across the three education categories. For word length there appears to be a slight trend for speakers in the lower education group to use shorter words. There does not appear to be any meaningful difference between the groups with respect to speech rate. Linear mixed effects models fit to the

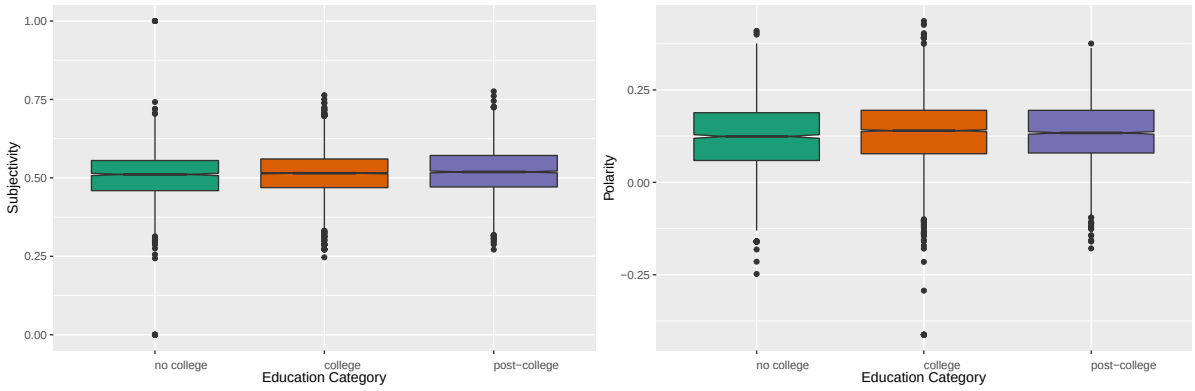

Figure 4.85: Segment subjectivity (left) and polarity (right) by education

data show a weak yet significant effect of education category on average word length ($\eta^2 = 0.029$, $p < 0.001$), but not on speech rate.

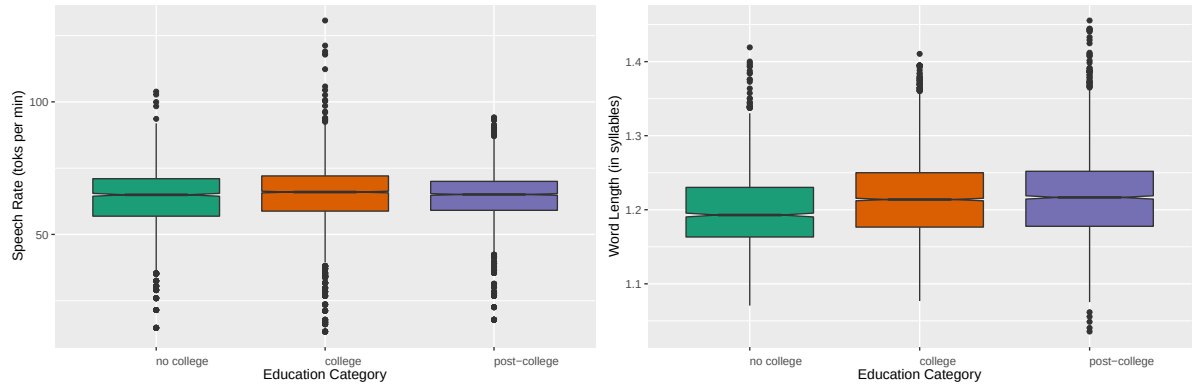

Figure 4.86: Speech rate (left) and word length (right) by education

4.5.4.9 Top informative ngrams

As detailed in chapter 3, the top 2000 ngrams (unigrams and bigrams combined) according to information gain for each sociodemographic trait of focus were extracted to serve as predictor features for the models presented in chapters 5 and 6.

Table 4.9 presents the top 20 ngrams according to information gain for speaker education. The full 2000 are not shown here for reasons of space.

Table 4.9: Top 20 informative ngrams for education

attributes	infogain
it_you	0.0223163
of	0.0192943
be_kind	0.0182232
kind	0.0156410
um_yeah	0.0149525
continue	0.0148577
kind_of	0.0138285
man_you	0.0137595
remember_i	0.0135244
responsibility	0.0130689
mail	0.0130483
grad	0.0127489
be_cheap	0.0124510
beautiful	0.0124386
careful	0.0123093
different_culture	0.0122196
i_friend	0.0119367
berkeley	0.0118617
week_i	0.0116626
sort	0.0116148

Though some of the top ngrams undoubtedly refer to education (e.g. "grad", "Berkeley") and some could be interpreted as being related to educational experience (e.g. "be_cheap", "different_culture"), the relationship between most of the top n-grams for education and the concept of education itself is unclear. It may be the case that the education data is so noisy that the top ngrams in this case are relatively uninformative.

Chapter 5: Baselines

In order to evaluate the performance of the multitask models discussed in chapter 6, it's first necessary to establish some baselines for comparison. Section 5.4 below first outlines naive baseline performance levels (majority class prediction) for each sociodemographic trait of focus. These naive baselines are then compared to predictions from several Single-Task Learning Multi-Layer Perceptron (STL-MLP) models trained on the NIST speaker corpus in section 5.5. The performance metrics of the STL-MLP models are then used as informed baselines against which the multi-task models detailed in chapter 6 are compared.

5.1 Preprocessing

Prior to any sort of evaluation, the data was preprocessed and split into training and testing sets. The preprocessing steps applied and a short explanation of each are listed below:

- **Centering**: The mean of each predictor column was subtracted from each value in that predictor column (resulting in a mean of 0 for each predictor).
- **Scaling**: Values in each predictor column were divided by the standard deviation of said predictor column (thereby normalizing variance among predictors).
- **Zero Variance Elimination**: Any predictor columns with zero variance (i.e. predictor columns with the same value for every observation) were removed.
- **K Nearest Neighbors Imputation**: In cases where a training example is missing data for a given predictor, the value of the missing data was estimated by averaging the values for that predictor from the k most similar training examples

in the data-set (K=5).

- **Yeo-Johnson Transformation**: A Yeo-Johnson power transformation (Yeo and Johnson, 2000) was applied in order to stabilize variance and coerce the predictors into more Gaussian-like distributions.
- **Correlation Filtering**: In order to eliminate unnecessary dimensionality, groups of predictors which were highly correlated with one another (r >= 0.8) were reduced by retaining only the predictor from that group with the lowest mean absolute correlation (i.e. the lowest average correlation with all other predictors).

All preprocessing steps were accomplished via the preprocessing functions available in the Caret R package (Kuhn et al., 2008).

5.2 Subsetting

After applying the preprocessing steps to the data as a whole, training and testing subsets were then created for each sociodemographic trait of focus. The subsets for each trait were balanced such that class representation for that trait in the data-set as a whole was roughly mirrored in both the training and testing subsets (e.g. as the data-set as a whole is roughly 16% Asian speakers, both training and test sets for ethnicity were balanced to be comprised of roughly 16% Asian speakers). For each trait, roughly 75% of the data was used for training and roughly 25% was held out for testing. The training and testing subsets were designed such that segments from the same speaker appeared in either the training or the testing subsets, but not in both. Segments in the corpus for which information was uncollected for a particular sociodemographic trait were not included in either the testing or training sets created for that trait. Each training subset included all acoustic, phonetic, and lexical features described in chapter 4, as well as those of the top 2000 most informative n-grams for that specific trait as determined by information gain which were not remove during correlation filtering.

5.3 Addressing Class Imbalance

As addressed in chapter 4, the NIST corpus is imbalanced for most of the sociodemographic traits of focus. Large class imbalances tend to be problematic for certain types of machine learning models, leading to over-prediction of the majority class and under prediction of the minority class(es). A common way of addressing a large class imbalance is to artificially re-sample the data such that the majority and minority classes are roughly equally represented in the training data. This can be done one of two ways– either one can remove examples of the majority class until equal class representation is reached (under-sampling), or one can add additional minority class examples until equal representation is reached (over-sampling). In practice, it is common to use a combination of both under-sampling and over-sampling when dealing with imbalanced data– over-sampling to boost the minority class numbers, and under-sampling around class boundaries in order to clean up the training set a bit.

To determine the relative benefit of resampling the training data for the present tasks, the informed baseline models reported below were trained both with and without resampling. The resampling algorithm used was an ensemble of the Synthetic Minority over-sampling Technique (SMOTE; Chawla et al., 2002) and Wilson's Edited Nearest Neighbor Rule (ENN; Wilson, 1972). SMOTE synthetically creates additional instances of the minority class(es) by interpolating between similar existing minority class instances while ENN removes instances whose class does not coincide with the majority class vote of the three nearest neighbor instances. See Batista et al. (2004) for a detailed description of this particular sampling combination and its benefits. In all cases, resampling either had no effect on model performance or slightly degraded model performance compared to the training runs without resampling. For this reason, in addition to the problems posed by using re-sampled data to train the multi-task nets

reported in chapter 6, only the performance metrics from training runs using non-re-sampled data are reported both below and in chapter 6.

5.4 Naive Baselines

The naive baseline used here is majority class prediction. Majority class prediction simply predicts each example in the test set to be a member of the majority class. For example, if a data set is comprised of 90% class A and 10% class B, majority class prediction will predict class A for every example, resulting in a baseline accuracy of 90%.

Tables 5.1 through 5.5 present the class distributions for the test-sets balanced for each trait.

Table 5.1: Class frequencies and proportions for sex in the testing subset

sex	freq	proportion
female	1390	0.55
male	1120	0.45

Table 5.2: Class frequencies and proportions for ethnicity in the testing subset

eth	freq	proportion
white	1380	0.63
black	380	0.17
asian	320	0.15
hispanic	120	0.05

Table 5.3: Class frequencies and proportions for age in the testing subset

age	freq	proportion
26-35	760	0.33
16-25	510	0.22
36-45	500	0.21
46-55	370	0.16
56+	190	0.08

Table 5.4: Class frequencies and proportions for region in the testing subset

reg	freq	proportion
northeast	860	0.39
south	500	0.23
midwest	430	0.20
west	410	0.19

Table 5.5: Class frequencies and proportions for education in the testing subset

edu	freq	proportion
college	1190	0.50
post-college	810	0.34
no_college	380	0.16

Table 5.6 presents the majority class baseline metrics for macro F1, weighted F1, and Accuracy for each sociodemographic trait of focus. Macro F1 refers to the unweighted average of F1 scores for each individual class, while weighted F1 refers to the average of F1 scores for each individual class weighted by class support. Macro F1 scores are more sensitive to error rates in minority classes, while weighted F1 scores give a better picture of model performance as a whole.

Table 5.6: Evaluation metrics for majority class baseline predictions

Trait	F1_macro	F1_weighted	Accuracy
Sex	0.356	0.395	0.554
Ethnicity	0.193	0.484	0.627
Age	0.098	0.160	0.326
Region	0.141	0.220	0.391
Education	0.222	0.333	0.500

5.5 Informed (Single-Task Learning) Baselines

For an informed baseline, Single-task multi-layer perceptron (STL-MLP) models were trained and evaluated on the trait-specific training/testing subsets described in section 5.2. In each case, the model consisted of an input layer, two dense, fully connected hidden layers using rectified-linear unit activation and dropout, and an output layer using soft-max activation over a number of neurons equal to the number of classes in the target trait. The optimization function used during training was macro-averaged F1 score. The basic architecture of the informed baseline models is shown in figure 5.1.

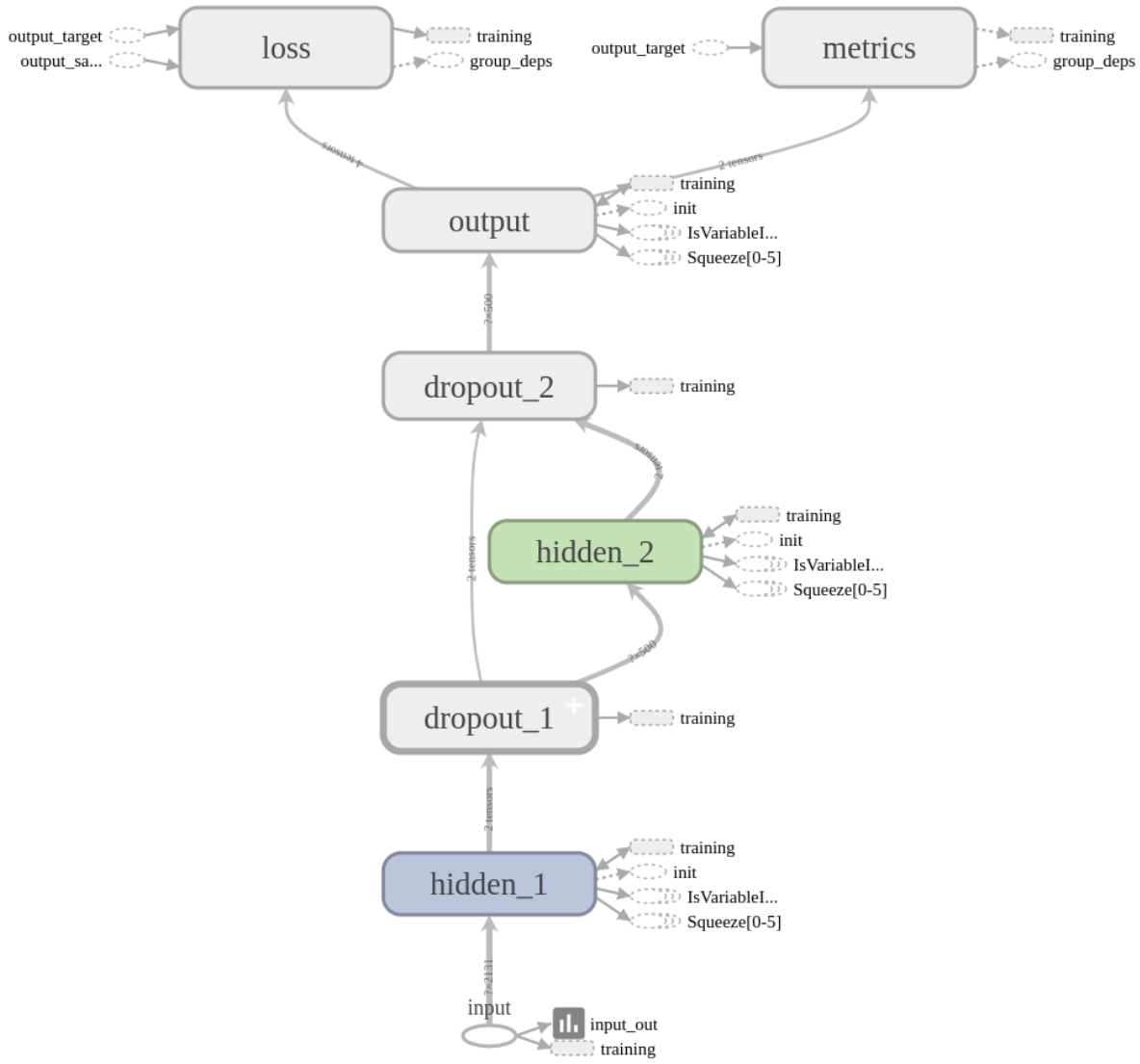

Figure 5.1: Basic model architecture for STL-MLP informed baselines

All informed baseline models were trained in Keras (Chollet, 2015) using TensorFlow (Abadi et al., 2016) as a back-end. Model hyper-parameters were tuned using the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011) implemented in Hyperopt (Bergstra et al., 2013), using k-fold cross validation and optimizing for macro-averaged F1. The hyper-parameters tuned and the distributions over which they were tuned are reported in table 5.7.

Table 5.7: Hyper-parameter search space

Hyper-Param	Search space
batch size	{32, 64, 128}
epochs	{10, 11, 12, ..., 100}
dropout rate	$\{x \epsilon \text{IR} \mid 0.1 \leq x \leq 0.9\}$
kernel initialization	{Glorot normal, Glorot uniform, He normal, He uniform}
optimizer	{adam, adadelta, rmsprop, sgd}
hidden_1 size	{10, 11, 12, ..., 650}
hidden_2 size	{10, 11, 12, ..., 650}

Final, tuned model parameters can be found in Appendix A.

To obtain a more accurate picture of model performance, a total of five models were trained using the final, tuned hyper-parameters for each trait of focus. All metrics for F1, Accuracy, and values in the confusion matrices reported below are the averaged results of those five training runs for each trait of focus.

5.5.1 Sex

The average macro F1 Score, weighted F1 score, and accuracy values for the five STL-MLP informed baselines trained to predict speaker sex are presented in table 5.8. The metrics for the naive, majority class baseline are included for comparison.

Table 5.8: Evaluaton metrics for sex baseline models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.356	0.395	0.554
STL_MLP	0.973	0.973	0.973

The informed baselines perform well above the naive baseline for all evaluation metrics. Figure 5.2 shows a confusion matrix comparing the true labels to the labels predicted by the informed baselines for the test set. In each cell, the gray text reports the average number of training samples belonging to that cell over the five informed baseline training runs, and the black text reports the percentage of samples in that particular row that are contained within the given cell. Percentages may be interpreted as the percent of true labels for a given class (represented by row) that received a given classification prediction (represented by column). Cell shading is computed via the percentage values.

The high model performance on predicting speaker sex is perhaps unsurprising given that sex had by far the most number of predictors that reached significance in appendix C. A two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten and Hinton, 2008) of the training and testing data is presented in figure 5.3. Each individual point on the plots represents one training or test example. Diamonds represent class means. The two classes are quite clearly distinct, and nearly linearly separable. It appears that the features extracted and selected for sex distinguish these two classes well.

5.5.2 Ethnicity

The average macro F1 Score, weighted F1 score, and accuracy values for the five STL-MLP models trained to predict speaker ethnicity are presented in table 5.9. The metrics

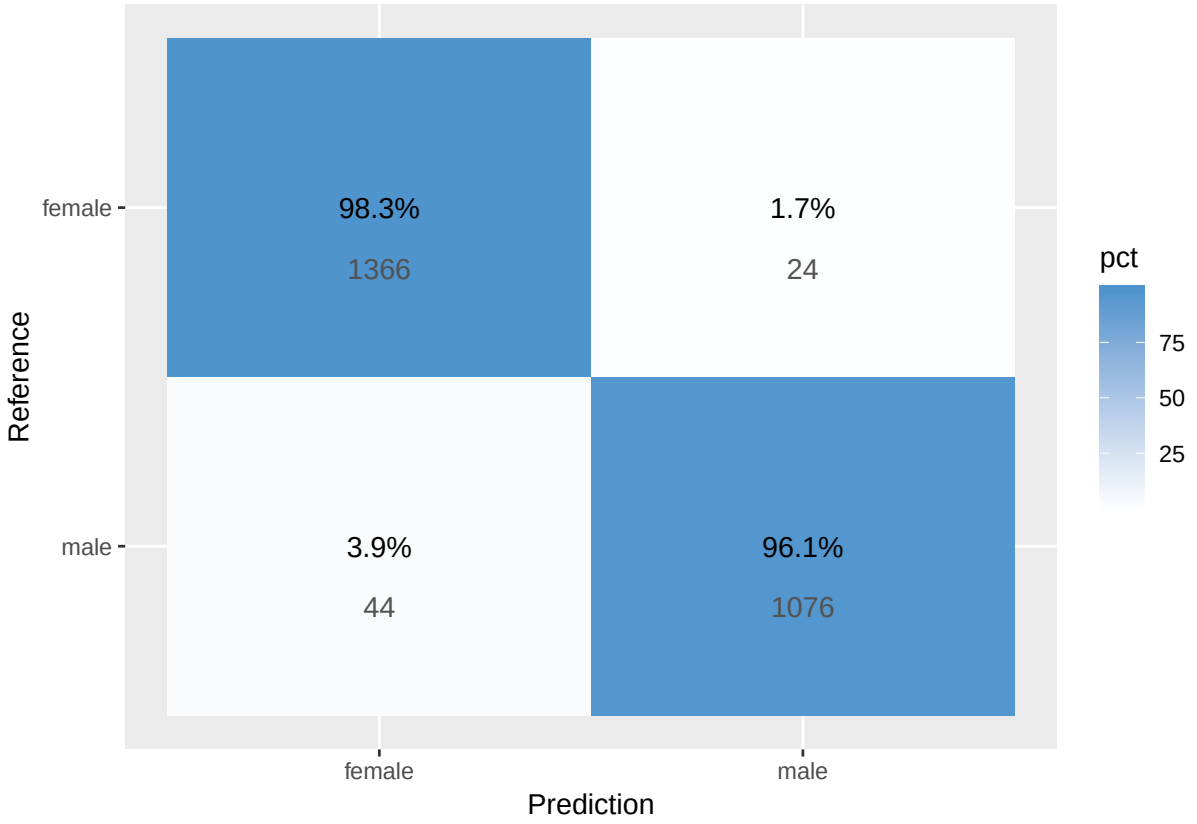

Figure 5.2: Confusion matrix for STL sex models.

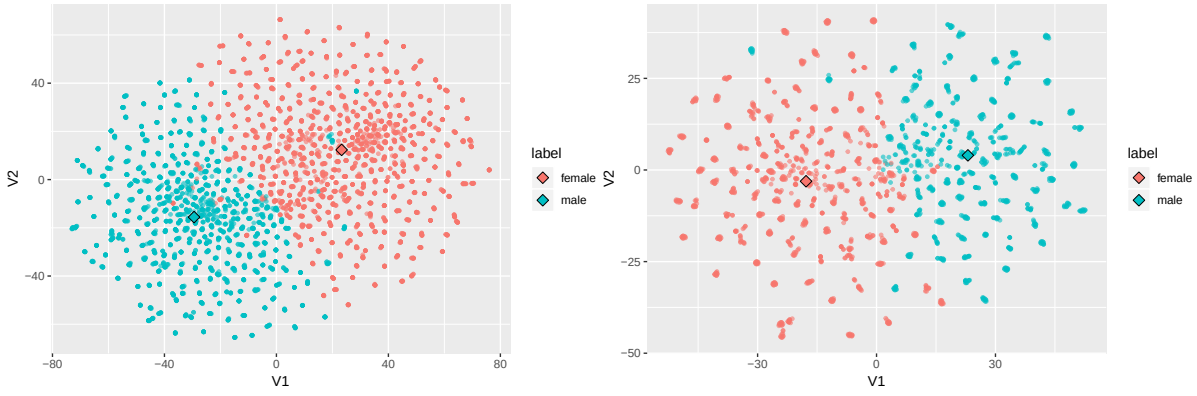

Figure 5.3: 2D t-SNE visualization of training (left) and test (right) data for sex.

for the naive, majority class baseline are included for comparison.

Table 5.9: Evaluaton metrics for ethnicity baseline models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.193	0.484	0.627
STL_MLP	0.790	0.879	0.873

As with speaker sex, the informed baselines for ethnicity well outperform the naive baseline for all evaluation metrics. Though the gains are not as high as those for sex, ethnicity is a four-way classification problem within this data-set and thus naturally a more difficult challenge to begin with. Regardless, performance of the STL-MLP informed baseline models is quite good. Figure 5.4 shows a confusion matrix comparing the reference labels to the predicted labels for the informed baseline models. As with sex, the numbers in each cell are averaged over the five training runs, and percentages are calculated row-wise.

The models appear to have developed quite accurate representations for White and Asian speakers, with correct identification rates of around 97% and 78% respectively. The models appear to have more difficulty correctly identifying Black and Hispanic speakers, though correct identification remains around 62-68% for these two groups. The models appear to have the most difficulty in correctly identifying Hispanic speakers. This is perhaps unsurprising, as Hispanic speakers were by far the most under-represented class in the data.

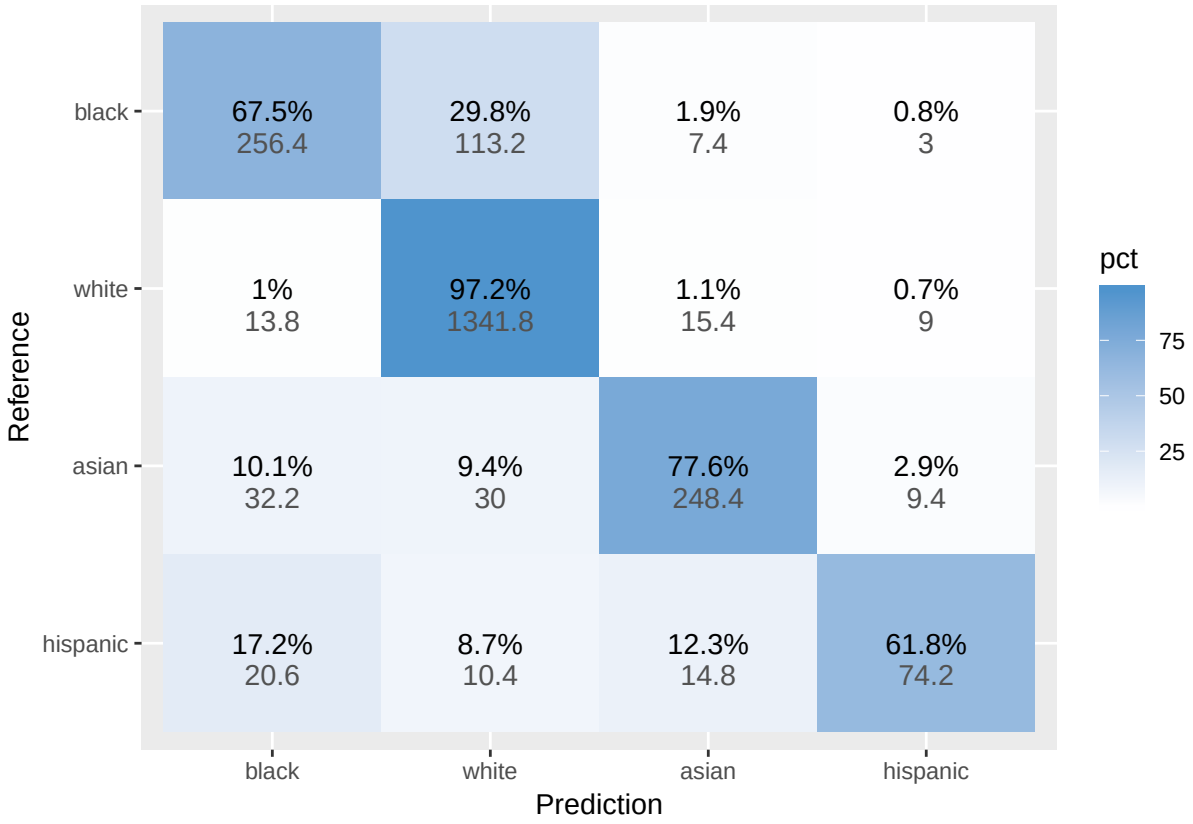

Figure 5.4: Confusion matrix for STL ethnicity models.

5.5.3 Age

The average macro F1 Score, weighted F1 score, and accuracy values for the five STL-MLP models trained to predict speaker age are presented in table 5.10. The metrics for the naive, majority class baseline are included for comparison.

Table 5.10: Evaluaton metrics for age baseline models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.098	0.160	0.326
STL_MLP	0.743	0.761	0.759

As with the other traits examined so far, the informed baselines for age outperform the naive baseline for all evaluation metrics. Figure 5.5 shows a confusion matrix comparing the reference labels to the predicted labels for the informed baseline models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training runs, and percentages are calculated row-wise.

For the most part, the model appears to identify age category for reference age group quite accurately, with a handful of mis-classifications spread out among the non-reference categories. There appears to be a tendency for misclassification either up or down one age category for most reference age groups, and a pattern for speakers in the 46-55 reference group to be misclassified as belonging to one of the next two youngest age groups.

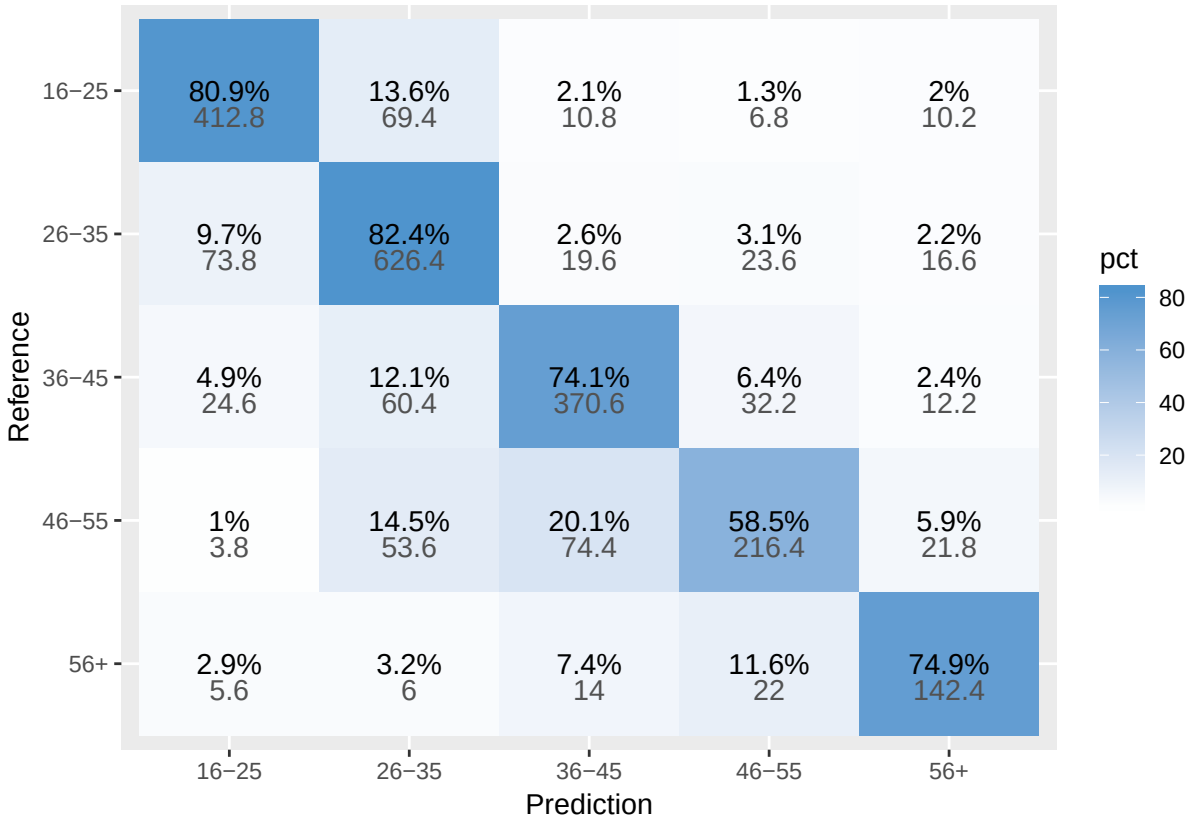

Figure 5.5: Confusion matrix for STL age models.

5.5.4 Region

The average macro F1 Score, weighted F1 score, and accuracy values for the five STL-MLP models trained to predict speaker region are presented in table 5.11. The metrics for the naive, majority class baseline are included for comparison.

Table 5.11: Evaluaton metrics for region baseline models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.141	0.22	0.391
STL_MLP	0.820	0.83	0.830

As with the other traits examined so far, the informed baselines for education far outperform the naive baseline for F1 and Accuracy. Figure 5.6 shows a confusion matrix comparing the reference labels to the predicted labels for the informed baseline models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training runs, and percentages are calculated row-wise.

Overall, performance of the informed baseline models on the region test data is quite good. Each reference group is correctly identified at least roughly 80% of the time, with, for the most part, just a handful of speakers from each group misclassified as belonging to a different region. There does however appear to be a slight tendency for over-prediction of the Northeast category.

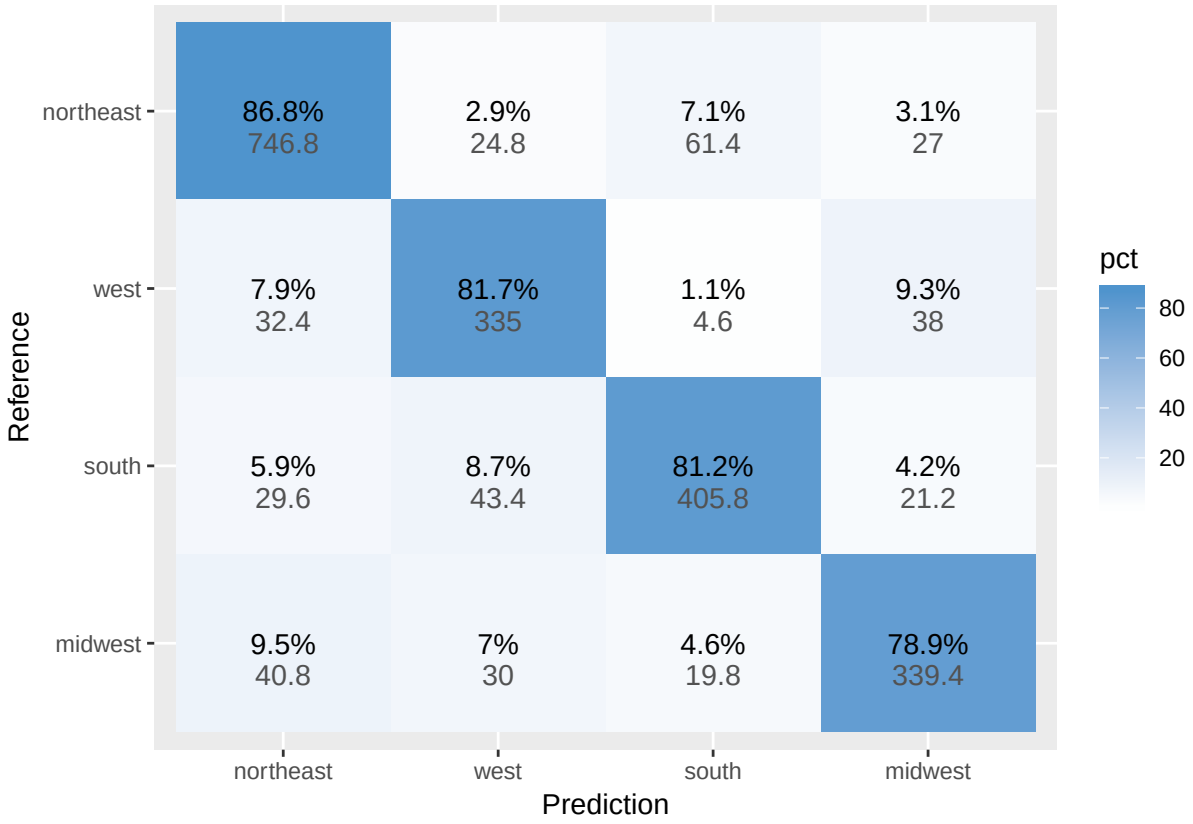

Figure 5.6: Confusion matrix for STL region models.

5.5.5 Education

The average macro F1 Score, weighted F1 score, and accuracy values for the five STL-MLP models trained to predict speaker education are presented in table 5.12. The metrics for the naive, majority class baseline are included for comparison.

Table 5.12: Evaluaton metrics for education baseline models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.222	0.333	0.500
STL_MLP	0.752	0.790	0.794

As with the other traits examined so far, the informed baselines for education far outperform the naive baseline for F1 and Accuracy. Figure 5.7 shows a confusion matrix comparing the reference labels to the predicted labels for the informed baseline models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training runs, and percentages are calculated row-wise.

As with the previous traits, each class for education is correctly identified with a reasonable degree of accuracy. Assuming the underlying trait of education years forms a continuum, one would expect the highest rates of misclassification to be off-by-one errors either up or down a class, and thus would expect classes representing the poles of that continuum to receive the lowest rates of misclassification, as these classes only have one neighboring class to be confused with as opposed to two. This is in fact the pattern observed for age in section 5.5.3. Interestingly however, correct classification for the middle 'college' education class is highest at around 89%, while correct classification of the pole classes is somewhat lower– roughly 70% for each. Even more interestingly, the most frequent misclassification error is misclassifying speakers in the highest education

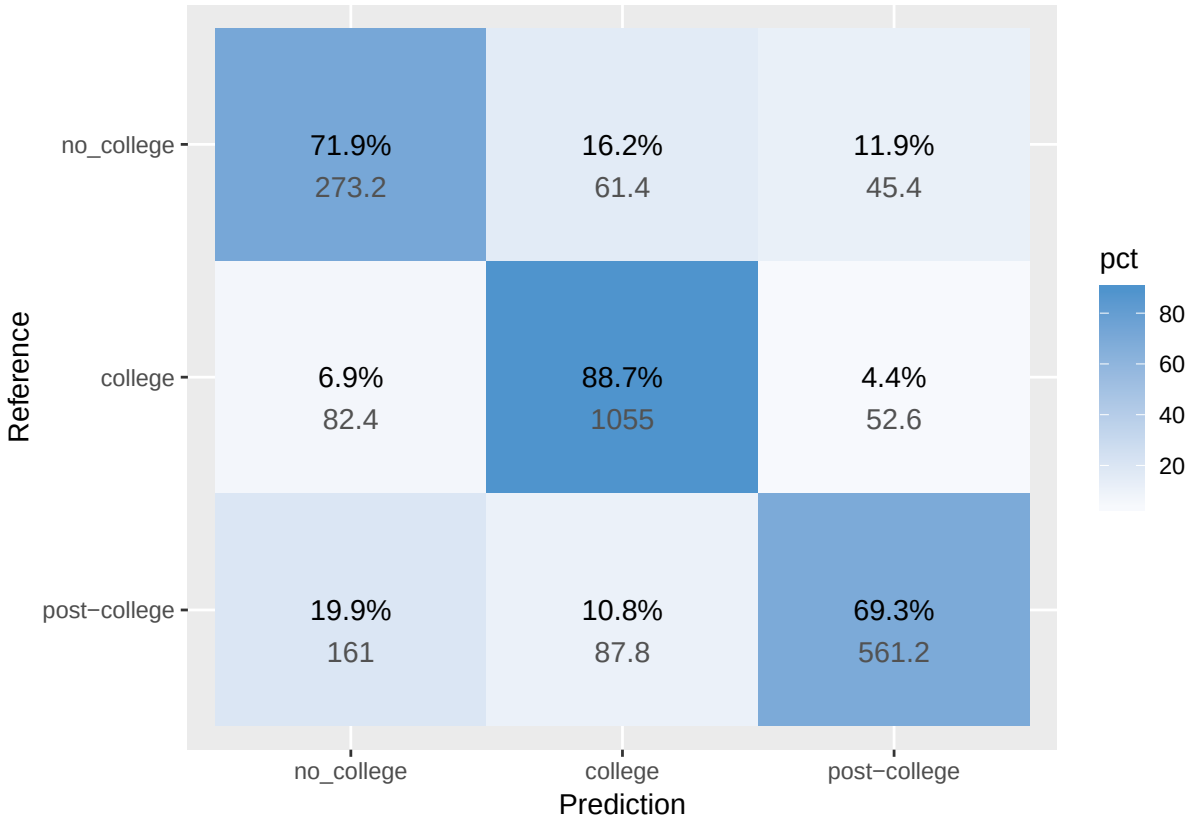

Figure 5.7: Confusion matrix for STL education models.

class as speakers belonging to the lowest education class. It is unclear what may be contributing to this particular misclassification pattern, though this may be in part due to the fact that the number of education years reported by speakers in the data-set may not always line up with the traditional education distinctions assumed in creating the three education experience buckets into which speakers have been placed for the purpose of this data-set.

5.6 Discussion

This chapter has established both naive and informed baselines against which to compare the Multi-Task Models which are the focus of this dissertation. These baselines also provide some insight into the informativeness of the extracted predictor features described in chapters 3 and 4 for each of the sociodemographic traits of focus. These features are particularly informative for predicting speaker sex, and the high performance of the models reported in section 5.5.1 leaves little room for improvement for the multi-task models. Performance of the informed baselines for ethnicity, age, education, and region is also quite high, though there is still quite a bit of room to grow for these tasks.

Overall, it appears that the features identified and discussed in chapter 4 perform reasonably well in distinguishing among the sociodemographic categories examined in this dissertation. An exploration of the relative contribution of each type of predictive feature to model performance for each sociodemographic trait is presented in chapter 7.

Chapter 6 reports performance of the MTL models and compares them to the baselines developed in this chapter. A detailed discussion comparing the STL and MTL models can be found in chapter 8.

Chapter 6: Results

The following sections present the basic structure and training procedure for the Multi-Task Learning Multi-Layer Perceptron (MTL-MLP) models, as well as evaluation metrics summarizing MTL model performance on the testing data. Performance of the MTL models on the test sets is compared to both naive baseline (majority class) model performance as well as informed baseline (STL-MLP) model performance on the same test sets.

6.1 Training and Testing Data

The data sets used to train and test the MTL-MLP models were the same data-sets used in chapter 5 to train and test the baseline models. Performance of all models discussed in this dissertation is therefore directly comparable. Refer to chapter 5 for a detailed description of the preprocessing and subsetting operations involved in generating these testing and training data sets.

6.2 Multi-Task Learning Model Description

For each trait of focus, the basic architecture of the MTL-MLP models was as follows. Each model was comprised of one shared, fully connected hidden layer using rectified-

linear unit activation and dropout, which then fed five separate (trait-specific) fully connected hidden layers also using rectified-linear activation and dropout. The trait-specific hidden layers each fed an individual output layer using soft-max activation over a number of neurons equal to the number of classes in the target trait for that specific output layer. A masking layer was applied directly prior to each trait-specific output layer, so as to eliminate error back-propagation from that output layer in cases where the class label for that particular trait was unknown for a given training example. Output layers for each individual trait were assigned a weight between 0.1 and 1.0, used in calculating overall loss during training. The optimization function used during training was macro-averaged F1 score. Each model took as inputs the training data as well as five masking tensors (one for each trait of focus), and had as outputs class probabilities for each of the five traits of focus.

One can conceive of the MTL model design somewhat like a tree with five separate branches. The model inputs and the shared hidden layer make up the 'trunk,' and each trait has its own 'branch' consisting of a hidden layer, a masking layer, and an output layer which are specific to that particular trait. The weight assigned to the output layer for each 'branch' determines the degree of importance that particular branch is given during model training. A diagram of the MTL network design used for the models reported in this chapter is shown in figure 6.1. The design in figure 6.1 only includes the first two 'branches' of the MTL models, sex and ethnicity, for reasons of space. The three branches left out of figure 6.1 are identical in structure to the sex and ethnicity branches, and all MTL models included all 5 branches.

All MTL-MLP models were trained in Keras (Chollet, 2015) using TensorFlow (Abadi et al., 2016) as a back-end. Model hyper-parameters were tuned using the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011) implemented in Hyperopt (Bergstra et al., 2013), using k-fold cross-validation and optimizing for macro-

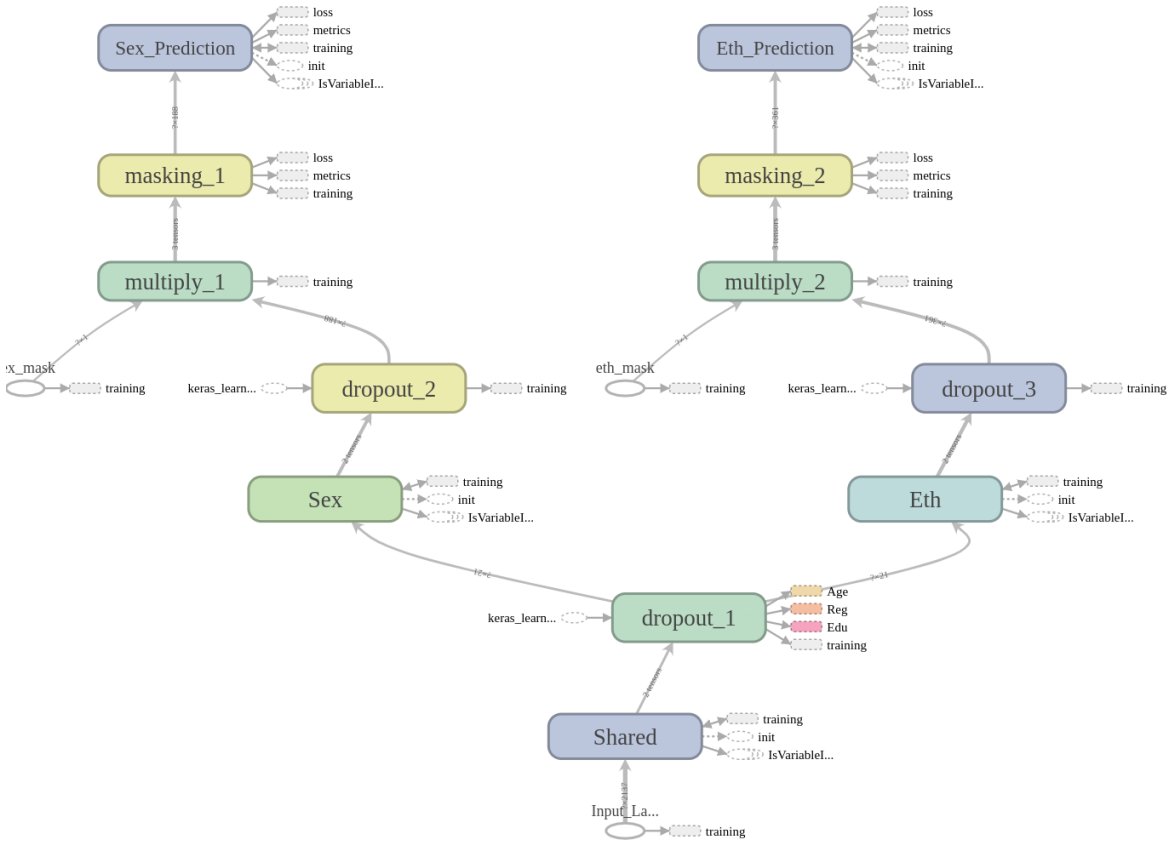

Figure 6.1: Basic model design for MTL-MLP models

averaged F1. The hyper-parameters tuned and the distributions over which they were tuned are reported in table 6.1. The final tuned hyper-parameters used in the models reported below for each trait of focus may be found in appendix B.

Table 6.1: Hyper-parameter search space for MTL-MLP models

Hyper-Param	Search space
batch size	{32, 64, 128}
epochs	{10, 11, 12, ..., 100}
dropout rate	$\{d \epsilon \text{IR} \mid 0.1 \leq d \leq 0.9\}$
kernel initialization	{Glorot normal, Glorot uniform, He normal, He uniform}
optimizer	{adam, adadelta, rmsprop, sgd}
shared layer size	{10, 11, 12, ..., 650}
trait-specific layer size	{10, 11, 12, ..., 650}
trait-specific loss weight	$\{w \epsilon \text{IR} \mid 0.1 \leq w \leq 1.0\}$

While each MTL model by design produced predictions for each of the five traits of focus in this dissertation, only predictions for the trait for which a particular model was specifically tuned are reported below. Models tuned to predict a particular trait used the testing and training subsets of the NIST corpus specific to that trait as described in section 5.2. For example, models tuned to predict ethnicity were hyper-parameter tuned, trained, and evaluated using the ethnicity training and testing subsets on which the ethnicity STL-MLPs were tuned, trained and evaluated, models tuned to predict sex were tuned, trained, and evaluated using the sex training and testing subsets on which the sex STL-MLPs were tuned, trained, and evaluated, and so on.

6.3 Multi-Task Learning Model Evaluation

The following sections report model evaluation metrics averaged over five independent training runs for each individual trait of focus.

6.3.1 Sex

The average macro F1 Score, weighted F1 score, and accuracy values for the five MTL-MLP models trained to predict speaker sex are presented in table 6.2. The metrics for the naive, majority class baseline as well as the informed, STL baseline are included for comparison.

Table 6.2: Evaluaton metrics for sex models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.356	0.395	0.554
STL_MLP	0.973	0.973	0.973
MTL_MLP	0.980	0.981	0.981

The MTL models appear on average to provide a slight boost over the STL models across the board. Figure 6.2 shows a confusion matrix comparing the true labels to the labels predicted by the MTL models for the test set. In each cell, the gray text reports the average number of training samples belonging to that cell over the five MTL training runs, and the black text reports the percentage of samples in that particular row that are contained within the given cell. Percentages may be interpreted as the percent of true labels for a given class (represented by row) that received a given classification prediction (represented by column). Cell shading is computed via the percentage values.

Comparing the confusion matrix for MTL predictions in figure 6.2 to the confu-

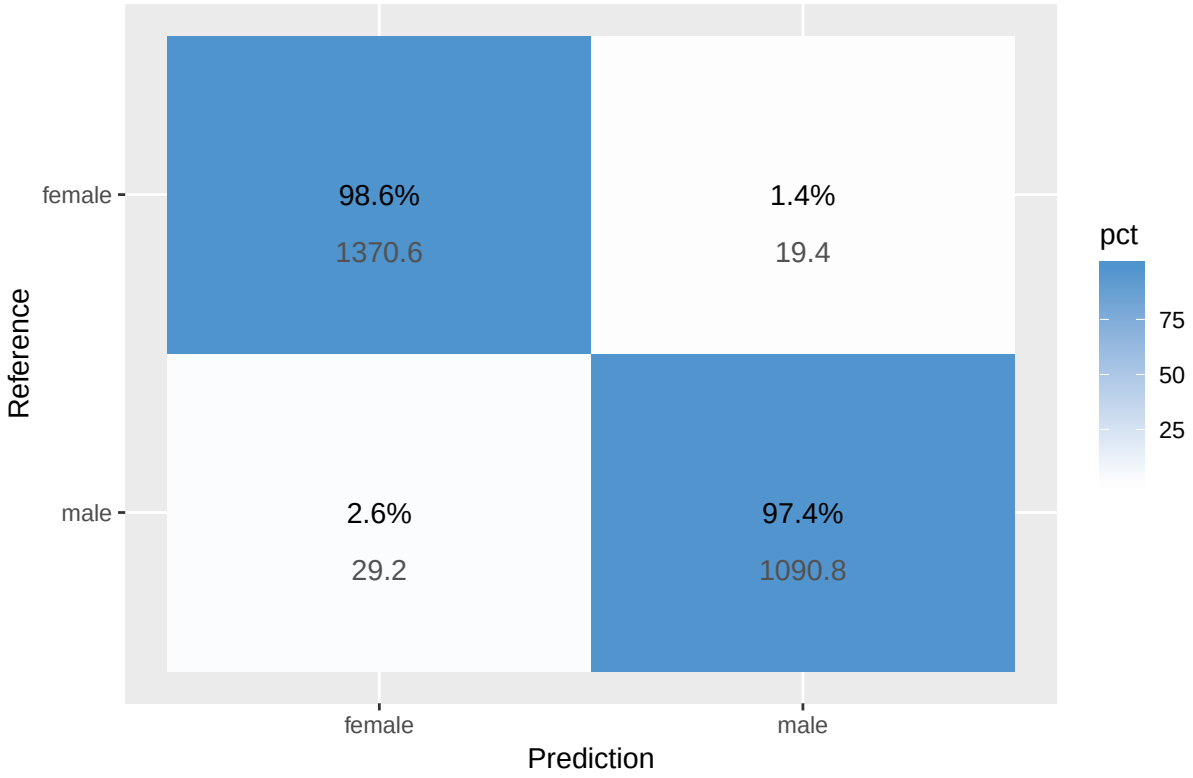

Figure 6.2: Averaged confusion matrix from the MTL training runs for sex.

sion matrix for STL predictions in figure 5.2, it seems the main area of improvement for the MTL models over the STL models is in properly classifying males. Incorrect identification of males as females dropped from 3.9% in the STL predictions to 2.6% in the MTL predictions. Improvement on classification error for females was more modest, dropping from 1.7% in the STL predictions to 1.4% in the MTL predictions.

6.3.2 Ethnicity

The average macro F1 Score, weighted F1 score, and accuracy values for the five MTL-MLP models trained to predict speaker ethnicity are presented in table 6.3. The metrics for the naive, majority class baseline as well as the informed, STL baseline are included for comparison.

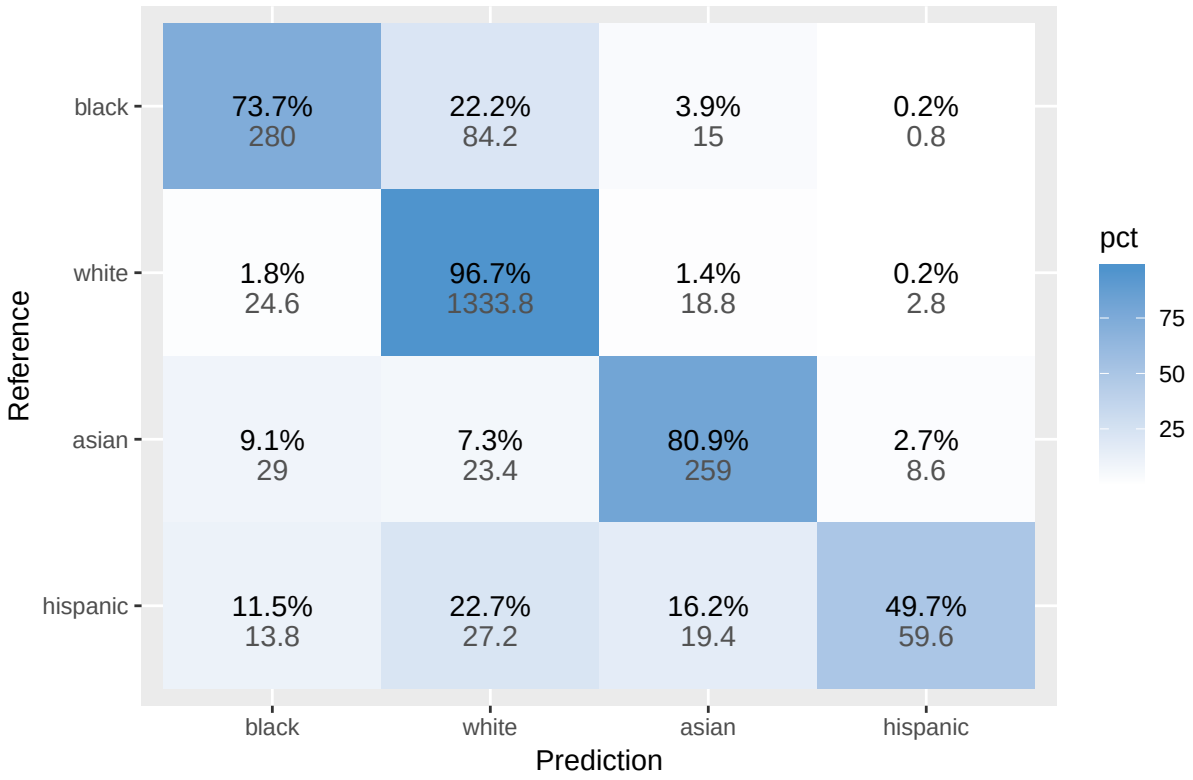

Figure 6.3: Averaged confusion matrix from the MTL training runs for ethnicity.

Table 6.3: Evaluaton metrics for ethnicity models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.193	0.484	0.627
STL_MLP	0.790	0.879	0.873
MTL_MLP	0.785	0.884	0.878

It appears that the MTL models for ethnicity slightly under-perform the STL models in terms of macro-F1 and slightly over-perform the STL models in terms of weighted F1 and Accuracy. Figure 6.3 shows a confusion matrix comparing the reference labels to the predicted labels for the MTL models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training runs, and percentages are calculated row-wise.

Comparing the confusion matrix for the MTL models in figure 6.3 to the confusion

matrix for the STL models in figure 5.4, it appears that the MTL models perform about the same as the STL models in identification of White speakers (0.5% error rate increase) but seriously under-performed the STL models in identification of Hispanic speakers (12.5% error rate increase). These two groups represent the extreme majority and minority classes, respectively. On the other hand, the MTL models outperformed the STL models in identification of the middle two classes, Black and Asian speakers, reasonably well– reducing error rates by 6.2% and 3.3% respectively. The gains in identification of Black and Asian speakers were sufficient to make up for the losses in identification of Hispanic speakers and give MTL models the edge in terms of weighted F1 and accuracy.

6.3.3 Age

The average macro F1 Score, weighted F1 score, and accuracy values for the five MTL-MLP models trained to predict speaker age are presented in table 6.4. The metrics for the naive, majority class baseline as well as the informed, STL baseline are included for comparison.

Table 6.4: Evaluaton metrics for age models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.098	0.160	0.326
STL_MLP	0.743	0.761	0.759
MTL_MLP	0.758	0.771	0.769

As with sex, the MTL models appear to have a slight advantage over the STL models across the board. Figure 6.4 shows a confusion matrix comparing the reference labels to the predicted labels for the MTL models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training

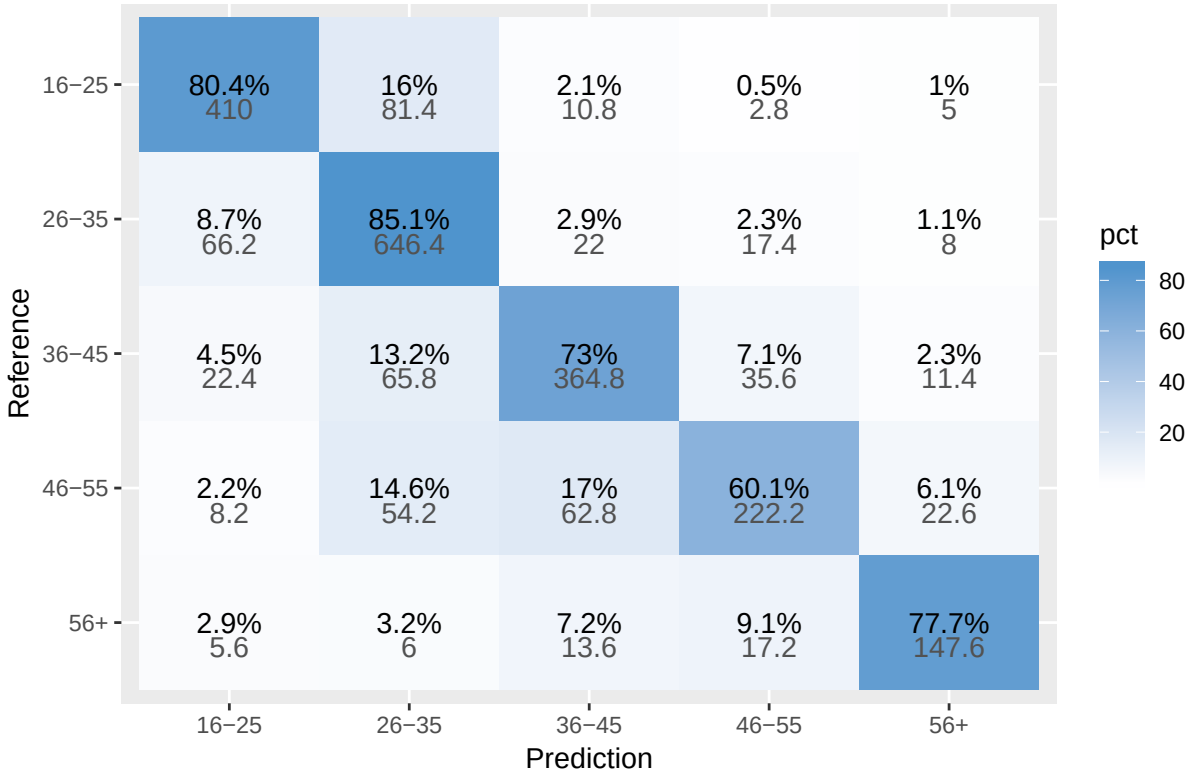

Figure 6.4: Averaged confusion matrix from the MTL training runs for age.

runs, and percentages are calculated row-wise.

Comparing the confusion matrix for the MTL predictions in figure 6.4 to the confusion matrix for STL predictions in figure 5.5, the main areas of improvement for MTL models over STL models are in classification of speakers age 56+ (the minority class) and speakers age 26-35 (the majority class), with roughly 2.8% and 2.7% error rate reductions, respectively. Identification error also went down in the MTL models compared to the STL models for speakers age 46-55; roughly a 1.5% reduction in error rate. Error rates for the other classes were roughly 0.5%-1% better for the STL models.

6.3.4 Region

The average macro F1 Score and accuracy values for the five MTL-MLP models trained to predict speaker region are presented in table 6.5. The metrics for the naive, majority class baseline as well as the informed, STL baseline are included for comparison.

Table 6.5: Evaluaton metrics for region models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.141	0.22	0.391
STL_MLP	0.820	0.83	0.830
MTL_MLP	0.818	0.83	0.829

For region, the STL and MTL models appear to be nearly identical in terms of evaluation metrics, with STL models having a minuscule edge in terms of macro F1 and overall accuracy. Figure 6.5 shows a confusion matrix comparing the reference labels to the predicted labels for the MTL models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training runs, and percentages are calculated row-wise.

In comparing the MTL confusion matrix in figure 6.5 with the STL confusion matrix in figure 5.6, it appears that the MTL models perform slightly better than the STL models on the majority class (Northeast, 0.8% error rate reduction) and the minority class (West, 2.4% error rate reduction). The STL models however outperform the MTL models by 4.6% on identification of the Midwest speakers.

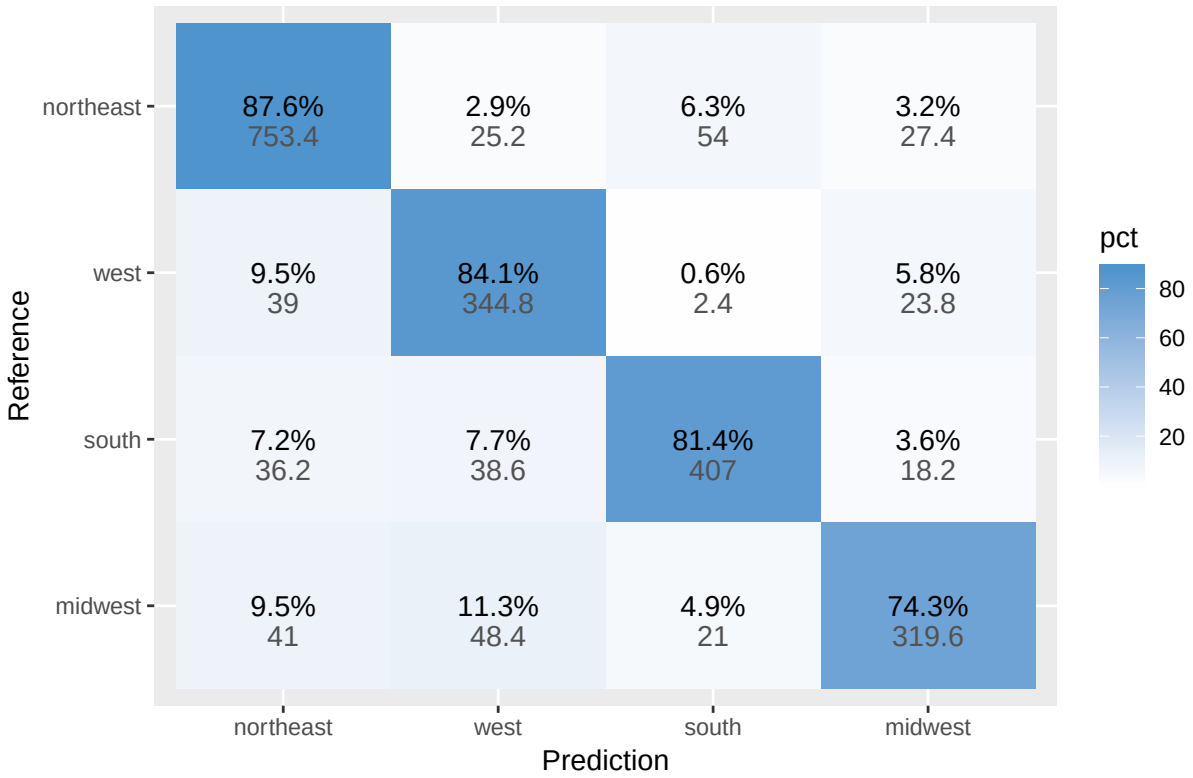

Figure 6.5: Averaged confusion matrix from the MTL training runs for region.

6.3.5 Education

The average macro F1 Score and accuracy values for the five MTL-MLP models trained to predict speaker education are presented in table 6.6. The metrics for the naive, majority class baseline as well as the informed, STL baseline are included for comparison.

Table 6.6: Evaluaton metrics for education models

Model	F1_macro	F1_weighted	Accuracy
MajClass	0.222	0.333	0.500
STL_MLP	0.752	0.790	0.794
MTL_MLP	0.758	0.800	0.802

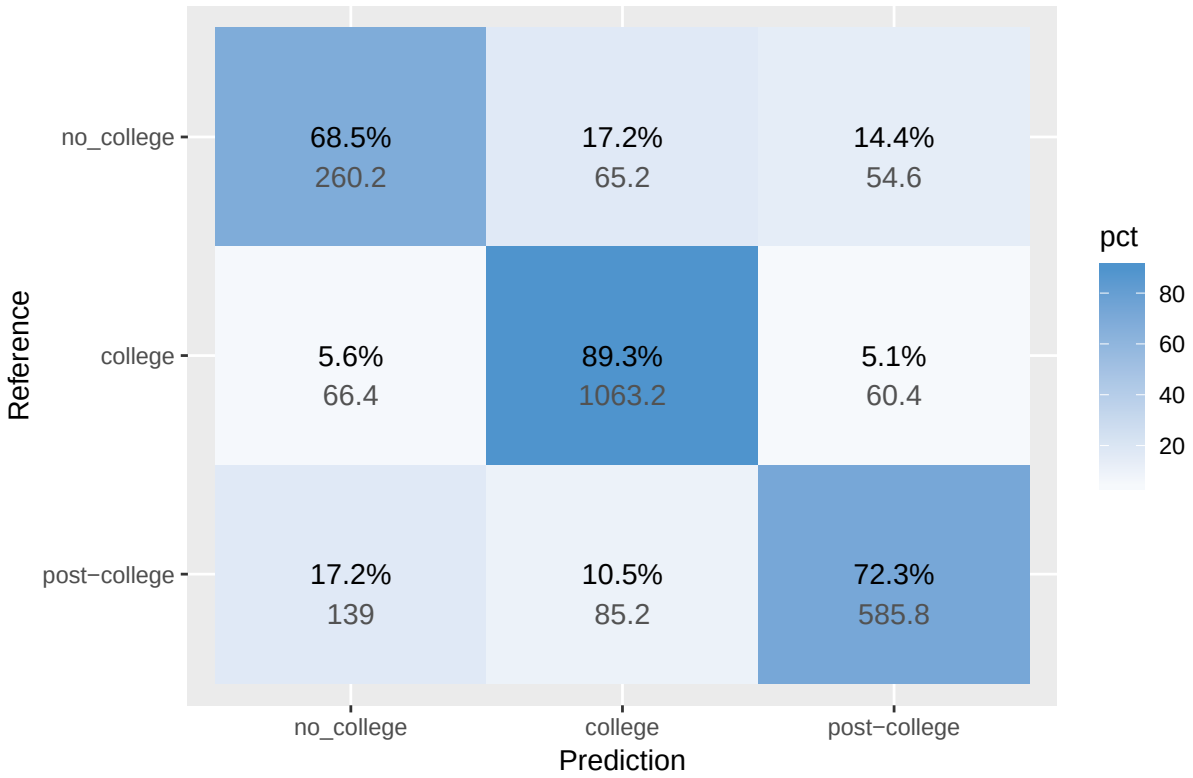

Figure 6.6: Averaged confusion matrix from the MTL training runs for education.

Though quite similar, the MTL models slightly outperform the STL models on all three evaluation metrics. Figure 6.6 shows a confusion matrix comparing the reference labels to the predicted labels for the MTL models. As with the previous traits, the number of testing examples that fall into each cell are averaged over the five training runs, and percentages are calculated row-wise.

In comparing the MTL confusion matrix in figure 6.6 with the STL confusion matrix in figure 5.7, it appears that the MTL models outperform the STL models on identification of the two most represented education groups in the corpus, with a 0.5% reduction of error rate in the majority class (college) and a 3% reduction in error rate for the next most represented class (post-college). The STL models outperformed the MTL models by roughly 3.4% in identification of the minority class (no-college).

6.4 Discussion

While the STL and MTL models appear to be quite close in terms of the selected evaluation metrics on the test sets, on the whole it does appear that in most cases using an MTL network design offered incremental improvement over an STL design. Interestingly there appears to be a slight tendency for MTL models to outperform STL models particularly in identification of well-represented (majority) classes. There may also be a very slight tendency to under-perform STL models on identification of the most extreme under-represented classes. This tendency and broader ramifications of the findings presented in this chapter are discussed in chapter 8.

Chapter 7: Feature Importance

Prior to discussing the differences between the performance of the STL and MTL network architectures, it is instructive to examine the relative weight that each type of model gave to each type of feature.

7.1 Measuring Feature Importance

The method for quantifying feature importance used in the following sections is an application of the permutation importance approach (Breiman, 2001; Fisher et al., 2018). The idea underlying permutation importance is that by randomly shuffling the data for one particular feature in the test set (thereby effectively scrambling the signal for that predictor while maintaining the original variance and mean) and measuring model performance before and after the shuffle, one can obtain a measurement for how strongly the model relies on that feature to make accurate predictions. In cases where a feature shuffle has a small impact on overall model performance, one can assume that the model does not rely much on that particular feature. In cases where shuffling the data for a particular feature severely degrades model performance, one can assume that the model relies heavily on that feature in making predictions.

The reference metric by which random permutation impact on model performance is measured in this chapter is macro F1. For each trait, macro F1 scores were extracted

from the predictions of the highest performing model for each of the STL and MTL architectures using the original testing data. For each feature, the testing data for that feature were then randomly permuted and predictions were run again. Permuted prediction scores for macro F1 ($F1_{perm}$) were then subtracted from reference prediction scores ($F1_{ref}$) and subsequently normalized in order to obtain the percentage of error rate increase (ERI) resulting from each feature permutation as shown in equation (7.1). The higher the ERI caused by random permutation of a given feature, the more important that feature was to obtaining accurate predictions in the test-set.

$$ERI = \frac{F1_{ref} - F1_{perm}}{F1_{ref}} * 100 \tag{7.1}$$

Error rate increase was calculated both on an individual basis as well as on a feature group basis. Group ERI scores were calculated by permuting all features belonging to a particular group in the testing data at once and then comparing predictions on the group-permuted data to the reference predictions in the same manner as shown in equation (7.1). The top 50 most important individual features per model as well as the overall importance of each feature group per model are presented in the following sections.

It should be noted that measuring feature importance in complex neural networks such as the models in question below is necessarily a fraught and not entirely straightforward task. Individual features may interact with one another in the internal layers of the models, and calculating individual feature importance by permuting one feature at a time may not take into account this sort of internal interaction. Likewise, in situations where two or more features are similarly important but substantially mutually redundant, although having at least one of these features present may be important to model predictions, calculating permutation scores on an individual basis will likely

lead to misleadingly low importance scores for all features in the group. Finally, as permutation importance is calculated below by applying pre-trained models to randomly permuted test data, the importance scores may be affected by random quirks of the test-set. In essence, all feature importance metrics reported in this chapter should be taken with a grain of salt. That said, measuring permutation importance for individual features and groups of features may still allow us to glean a general overview of what types of features are important for which types of models, and may help to elucidate some of the differences between the STL and MTL architectures.

7.2 Sex

7.2.1 Individual feature importance

Figures 7.1 and 7.2 present a comparison of the top 50 most important features for the best performing STL and MTL models, respectively, trained to predict speaker sex. Feature group is color-coded: lexical features are presented in grey, phonetic features in red, and acoustic features in turquoise.

Unsurprisingly, several of the acoustic features examined (mean pitch, jitter, shimmer, HNR) are near the top for both model architectures. For the MTL model, mean pitch is by far the most important feature, with random permutation of mean pitch degrading model performance by roughly 0.6%. A number of phonetic features make it into the top 50 as well, including both summary statistics of the vowel space as a whole (e.g. vowel space area, mean F1) and measurements of specific vowels at specific trajectory points (e.g. onset of EY along F2, off-glide of IY along F2). The acoustic and phonetic features included in the top 50 features in figures 7.1 and 7.2 generally line up with expectations and the observations made in chapter 4.

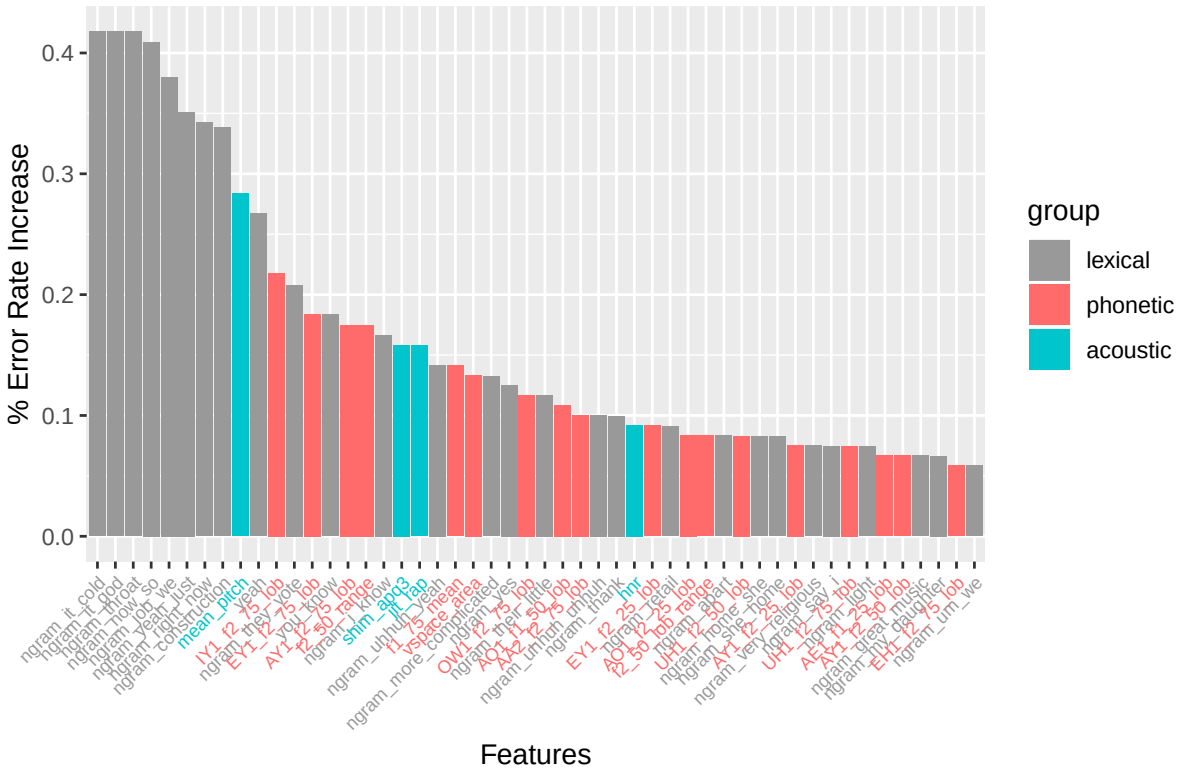

Figure 7.1: Top 50 individual sex features for STL architecture

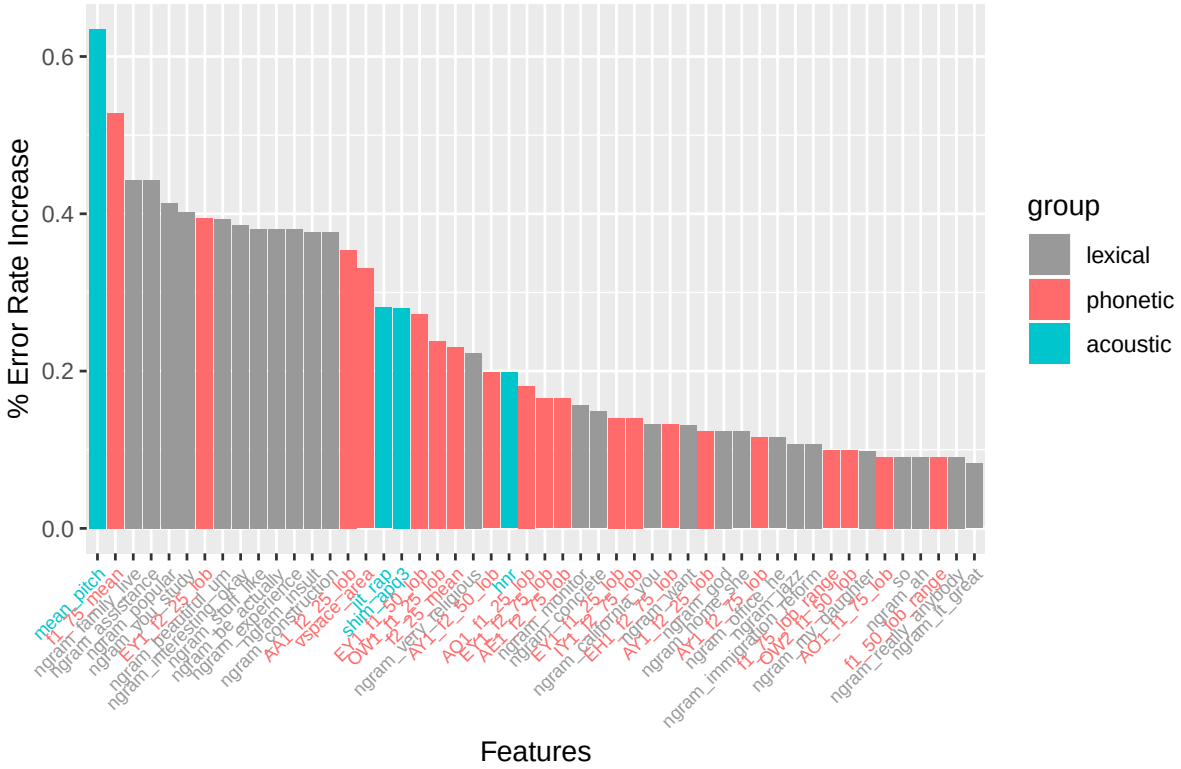

Figure 7.2: Top 50 individual sex features for MTL architecture

The lexical features that make it into the top 50 for each model are a bit more surprising.[1] First, the fact that the top 8 features for the STL model are lexical is contrary to expectations. Given the clear differences between male and female speakers for acoustic and phonetic variables shown in chapter 4, especially mean pitch, one would have expected more model reliance on these features than on any individual lexical feature. Second, the types of lexical features that are most important in these models do not line up with the types of lexical features that were found to be most important according to information gain for distinguishing between the sexes in chapter 4. Nearly all of the top lexical features according to information gain in chapter 4 included either a gendered kinship term (e.g. "husband", "wife", "daughter"), a discourse marker (e.g. "like", "oh", "yeah") or a pause filler (e.g. "uh", "um", "uhhuh"). While many of the lexical features in the top 50 do belong to these groups (e.g. "stuff like", "beautiful um", "my daughter"), several of them also consist of specific content words (e.g. "throat", "immigration reform", "concrete"). It's possible that the high importance placed on specific content words in these models may be a result of covert over-fitting– that is, over-fitting not necessarily to the training set specifically, but rather to this corpus as a whole.[2] While apparently specific lexical items such as these were important predictors both in the training and test sets (otherwise the models would not have paid attention to them in training and they would not have such a high impact in testing), these features are unlikely to generalize to other corpora.

[1]Though it should perhaps be noted again here that, as mentioned in chapter 2, while the focus of most of the signal processing and sociolinguistic work on sex identification/differentiation has been on acoustic and phonetic features, a wide range of studies attempting to identify the sex of authors of textual data has shown that lexical features can be highly reliable for this task (e.g. Rao et al., 2010).

[2]This type of covert over-fitting phenomenon has been recently documented for lexical features used in other computational linguistics tasks as well. See for example Moosavi and Strube (2017).

Table 7.1: Feature group importance: Sex STL

group	group error increase	mean individual error increase
lexical	20.03%	0.002%
phonetic	2.94%	0.009%
acoustic	0.87%	0.097%

Table 7.2: Feature group importance: Sex MTL

group	group error increase	mean individual error increase
lexical	27.4%	0.009%
phonetic	2.9%	0.054%
acoustic	1.24%	0.202%

7.2.2 Feature type importance

Tables 7.1 and 7.2 present the overall permutation importance for each feature group calculated for macro F1, as well as the mean of the individual ERI scores for each feature belonging to that group.

At first glance, both architectures appear to place much more importance on lexical features than phonetic or acoustic features. This is somewhat misleading however, as there were nearly ten times as many lexical features included in the data-set as there were phonetic and acoustic features combined. The mean macro f1 scores for individual features of each type make clear that, though as a whole lexical features were enormously important to model predictions, lexical features on average were less important individually to the model than acoustic and phonetic features.

Interestingly, it appears that the STL model is a bit more robust to poor signal quality within a given feature type than the MTL model overall. While permuting the phonetic features degraded model performance roughly equivalently in both STL and MTL models, permuting the Lexical and Acoustic features degraded model performance

substantially more for the MTL model than the STL model.

7.3 Ethnicity

7.3.1 Individual feature importance

Figures 7.3 and 7.4 present a comparison of the top 50 most important features for the best performing STL and MTL models, respectively, trained to predict speaker ethnicity. Feature group is color-coded: lexical features are presented in grey, phonetic features in red, and acoustic features in turquoise.

The most noticeable difference between the STL and MTL models in terms of feature importance is that the vast majority of the 50 most important features for the MTL model are lexical, whereas the set of top 50 features for the STL model has many more phonetic and acoustic features mixed in. This difference is somewhat difficult to interpret. It may have to do with the interaction between ethnicity and other sociodemographic traits and the fact that the MTL model is simultaneously developing representations of all traits whereas the STL model is singularly developing representations of ethnicity. For instance, the ethnic categories in both the training and testing sets are imbalanced with respect to sex– the two largest ethnic classes (White and Black) are predominantly female, whereas the two least frequent ethnic classes (Asian and Hispanic) are roughly evenly split between male and female. It's reasonable therefore to assume that some representation of speaker sex may be useful in predicting speaker ethnicity. The STL model may be relying directly on features such as pitch, HNR and jitter to model the relationship between sex and ethnicity whereas the MTL model may have developed other, more complex relationships within the shared layers that indicate speaker sex but which are more resilient to the presence or absence of any

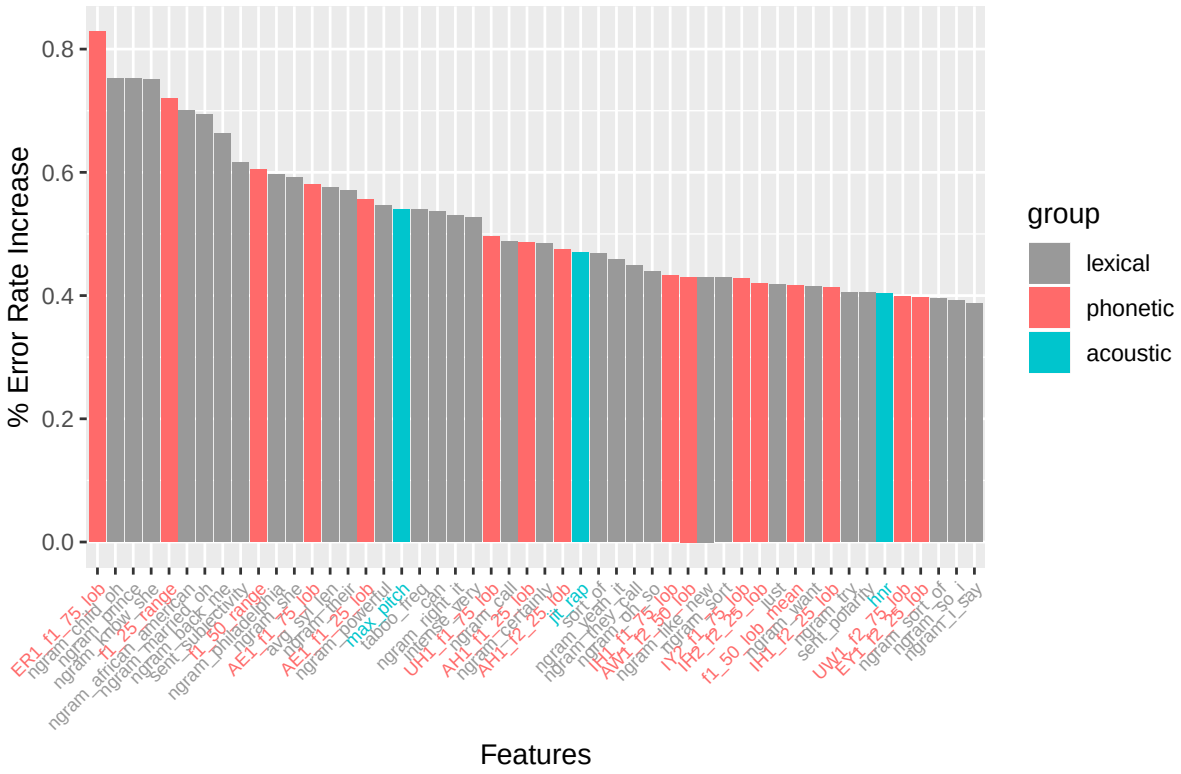

Figure 7.3: Top 50 individual ethnicity features for STL architecture

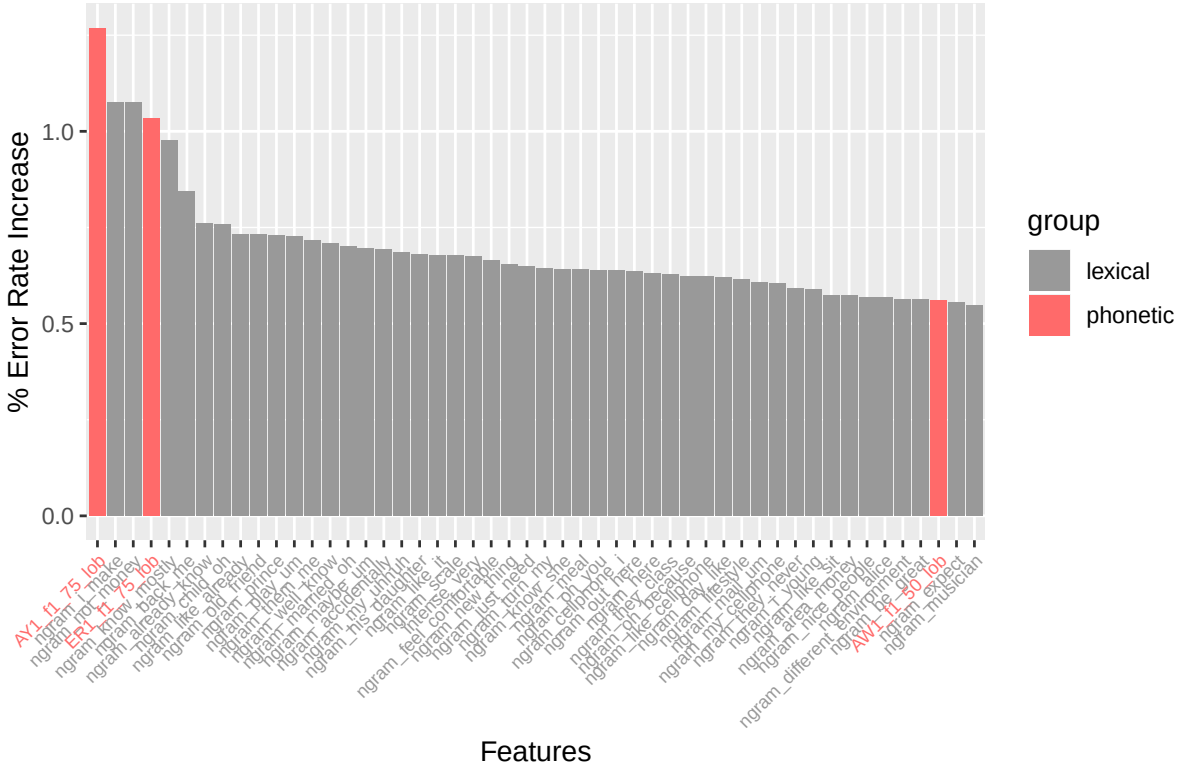

Figure 7.4: Top 50 individual ethnicity features for MTL architecture

Table 7.3: Feature group importance: Ethnicity STL

group	group error increase	mean individual error increase
lexical	78.33%	0.019%
acoustic	0.7%	0.336%
phonetic	0.25%	0.216%

Table 7.4: Feature group importance: Ethnicity MTL

group	group error increase	mean individual error increase
lexical	66.95%	0.036%
phonetic	5.47%	-0.045%
acoustic	0.27%	0.058%

individual acoustic or phonetic feature.

7.3.2 Feature type importance

Tables 7.3 and 7.4 present the overall permutation importance for each feature group calculated for Macro F1, as well as the mean of the individual ERI scores for each feature belonging to that group.

Both architectures appear to rely much more heavily on lexical features for ethnicity prediction relative to sex prediction, unsurprisingly. Interestingly, there appears to be a bit of a trade-off between reliance on lexical vs. phonetic features between the two architecture types, with MTL relying a bit more on the phonetic feature group and quite a bit less on the lexical feature group than the STL model. This is contrary to the distribution of feature groups seen in figures 7.3 and 7.4, and indicates that we should be cautious in drawing conclusions about overall model behavior from examining individual feature importance from just the top 50 most important features alone. It appears that, while more phonetic features individually make it into the top 50 for the

STL model than the MTL model, phonetic features as a group are more than 5 times more important in the MTL model than in the STL model.

Oddly, though the group importance of phonetic features is relatively substantial for the MTL model, the mean macro F1 score for phonetic features individually is actually negative. Negative importance scores indicate that the random permutation helped rather than hindered model performance– in other words, a negative importance score indicates that a feature is worse than random noise in terms of test-set predictions. The fact that while phonetic features individually were on average slightly worse than random noise yet as a block contributed to a 5.47% reduction in error rate in the MTL model is consistent with the notion developed above that the predictive power of these features in the MTL model stems from inter-feature interactions within the shared hidden layer which are relatively resilient to the removal of any one particular feature. This also highlights the importance of calculating feature group importance by permuting all features within that group simultaneously and calculating ERI based on the resulting predictions rather than simply adding together the individual importance scores for each feature within the group. Summing the ERI scores for each individual phonetic feature for the MTL ethnicity model results in a net ERI score of -5.94%, which would erroneously indicate that replacing all phonetic features with random noise would *increase* model performance by 5.94%. What we actually find is completely the opposite: replacing all phonetic features with random noise actually *decreases* model performance by 5.47%.

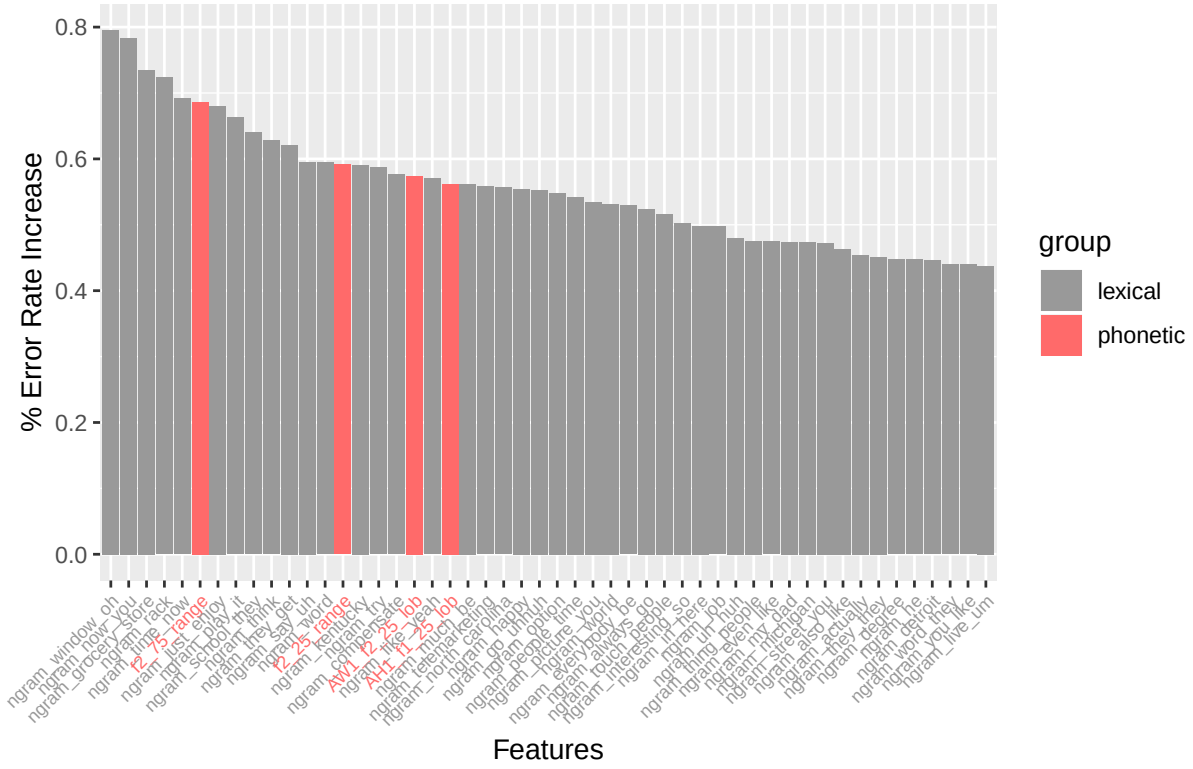

Figure 7.5: Top 50 individual age features for STL architecture

7.4 Age

7.4.1 Individual feature importance

Figures 7.5 and 7.6 present a comparison of the top 50 most important features for the best performing STL and MTL models, respectively, trained to predict speaker age. Feature group is color-coded: lexical features are presented in grey, phonetic features in red, and acoustic features in turquoise.

As with ethnicity, the models trained to predict speaker age appear to rely predominantly on lexical features, with a handful of acoustic and phonetic features cracking the top 50. In chapter 4 we saw that those ngrams which had the highest information gain for age tended to be those including either discourse markers or slang terms that

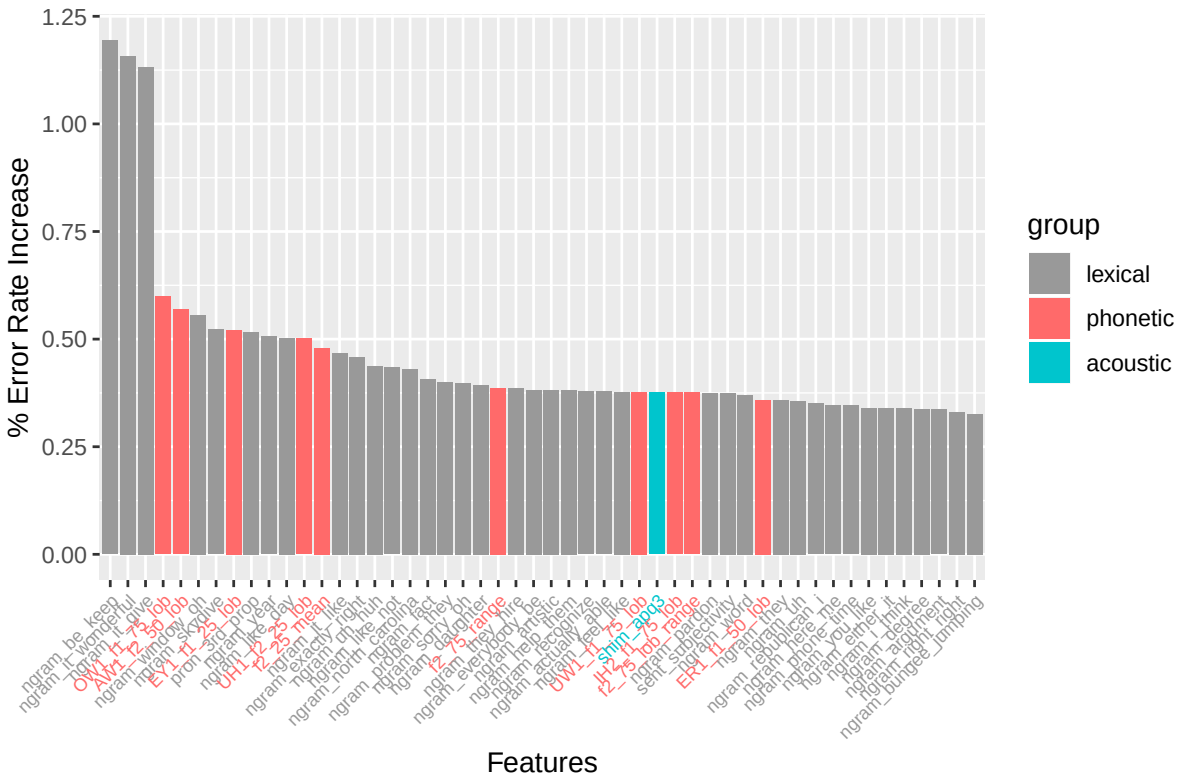

Figure 7.6: Top 50 individual age features for MTL architecture

showed a high degree of age-related differential usage (e.g. "yeah", "like", "cool") or terms that may be indicative of life-stage (e.g. "daughter", "wife"). The lexical features in the top fifty for the age models do include quite a few ngrams of this sort (e.g. "my dad", "job", "also like" for STL; "it like", "like not", "daughter" for MTL).

As with those models trained to predict sex however, many of the lexical features in the top fifty for model importance also contain specific content words that one would not expect given the information gain findings from chapter 4. The top 3 features for the MTL model for instance– each nearly twice as important as the next most important feature– include none of the sorts of lexical items one would expect. Oddly, a number of the top fifty features for the STL model are ngrams containing toponyms ("kentucky", "north carolina", "michigan", "detroit"). These toponyms all belong to the South and Midwest regions according to the coding schema used in this dissertation, which also happen to be the regions that skew the oldest. Recall from chapter 4 that there are far

Table 7.5: Feature group importance: Age STL

group	group error increase	mean individual error increase
lexical	71.73%	0.034%
phonetic	1.02%	0.029%
acoustic	0.04%	0.058%

Table 7.6: Feature group importance: Age MTL

group	group error increase	mean individual error increase
lexical	71.96%	0.022%
phonetic	0.43%	0.064%
acoustic	-0.2%	-0.062%

more younger speakers than older speakers in the data-set, and from chapters 5 and 6 that both STL and MTL models had more difficulty correctly identifying speakers belonging to the older age groups than those belonging to the younger age groups. It's possible that both the STL and MTL models trained to predict speaker age are to some degree reliant on a representation of speaker region to identify speakers belonging to older age groups, thereby taking advantage of the imbalance in the data-set between regions with respect to age, but that the MTL model does this indirectly via features jointly developed by the region and age output layers in the shared hidden layer during training whereas the STL model needs to do this more directly by relying on toponyms since it doesn't have any other way to represent the relation between region and age.

7.4.2 Feature type importance

Tables 7.5 and 7.6 present the overall permutation importance for each feature group calculated for Macro F1, as well as the mean of the individual ERI scores for each feature belonging to that group.

As with ethnicity, both architectures are heavily reliant on lexical features. Permuting the lexical features reduces the macro F1 prediction score by roughly 72% in both cases. In contrast, permuting phonetic features impacts the macro F1 prediction score by at most roughly 1%, and permuting the acoustic features has essentially no effect in the STL model and actually improves performance in the MTL model by 0.2%. It seems clear that, with a few phonetic feature exceptions, the high performance of the STL and MTL models on age prediction seen in chapters 5 and 6 are almost entirely driven by lexical features.

7.5 Region

7.5.1 Individual feature importance

Figures 7.7 and 7.8 present a comparison of the top 50 most important features for the best performing STL and MTL models, respectively, trained to predict speaker region. Feature group is color-coded: lexical features are presented in grey, phonetic features in red, and acoustic features in turquoise.

The models trained to predict speaker region appear to rely predominantly on lexical features, with just a couple of phonetic features making it into the top 50. As expected, acoustic features appear to have little impact on the region models.

Interestingly, though the top ngrams according to information gain for region from chapter 4 were, predictably, heavily centered around toponyms, relatively few toponyms make it into the top 50 most important features for either model. One explanation for this may lie in the nature of the metrics in question. Information Gain is at heart a measure of purity within groups delineated by a particular splitting factor. For instance, nearly every mention of "bay area" in the corpus belongs to a speaker from the Western

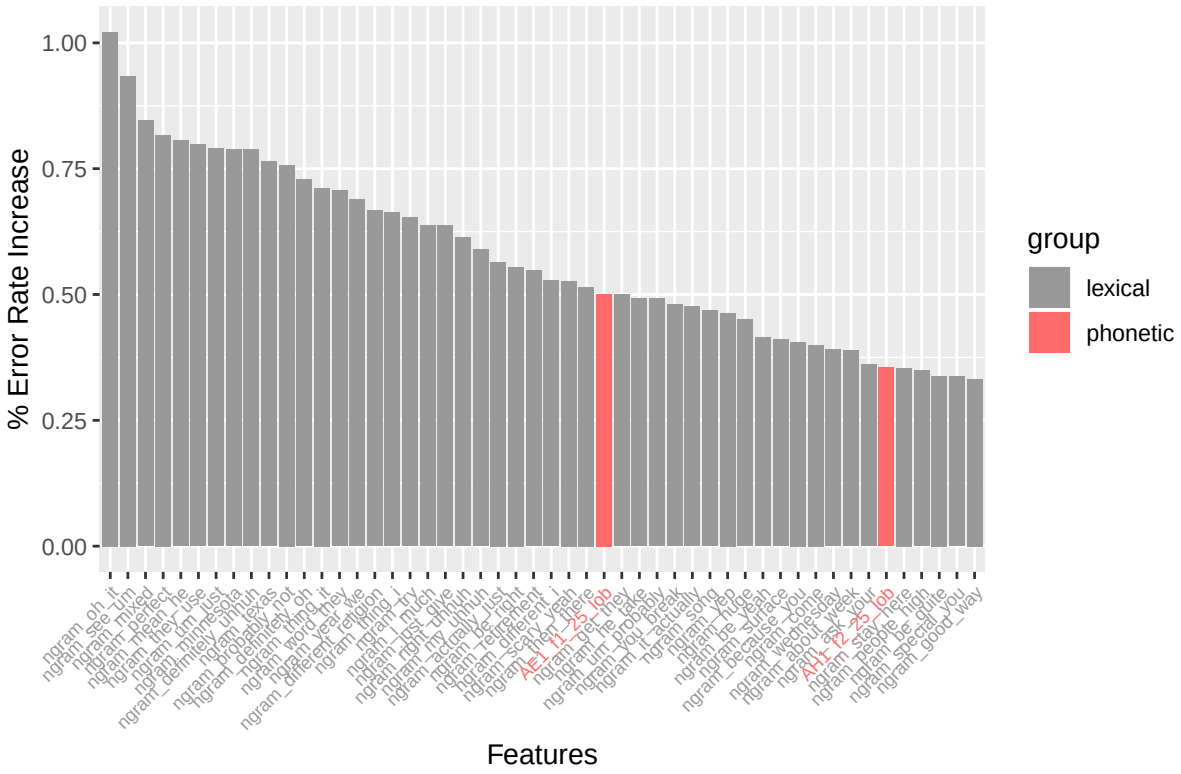

Figure 7.7: Top 50 individual region features for STL architecture

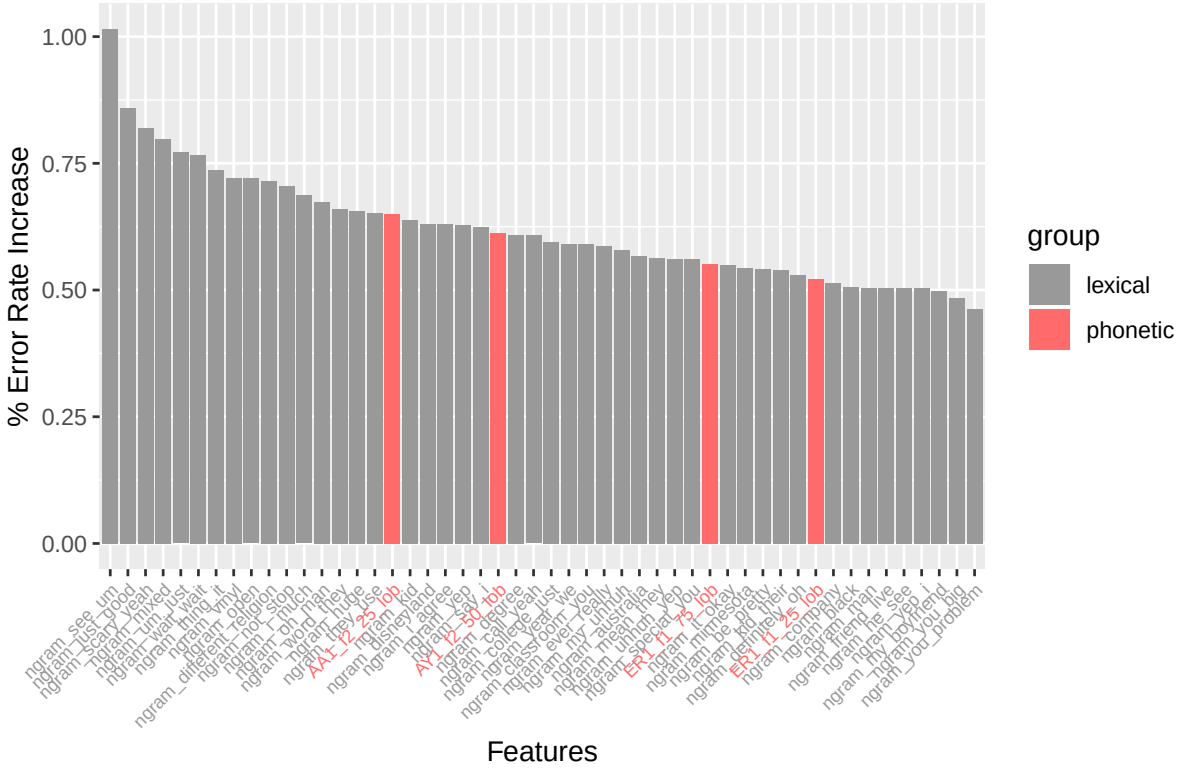

Figure 7.8: Top 50 individual region features for MTL architecture

Table 7.7: Feature group importance: Region STL

group	group error increase	mean individual error increase
lexical	68.76%	0.035%
phonetic	4.13%	-0.032%
acoustic	0.39%	0.021%

Table 7.8: Feature group importance: Region MTL

group	group error increase	mean individual error increase
lexical	67.36%	0.049%
phonetic	2.54%	0.083%
acoustic	0.02%	0.007%

region. Splitting the data based on mention of "bay area" therefore leads to a very pure sub-sample of speakers form the west, and thus high information gain. "Bay area" however is a relatively rare term used by few speakers, and the category of western speakers is a minority regional class in the data-set, so the actual predictive power of this particular feature is relatively low. On the whole, the lexical features with the highest ERI scores in both models seem to have a tendency to include discourse markers and pause fillers (e.g. "oh man", "see um", "yep", "yeah"). That the models appear to be paying more attention to the distribution of discourse markers and pause fillers rather than to specific toponyms is likely a good thing in terms of model generalization, as these items are far more frequent than toponyms in most genres of communication.

7.5.2 Feature type importance

Tables 7.7 and 7.8 present the overall permutation importance for each feature group calculated for Macro F1, as well as the mean of the individual ERI scores for each feature belonging to that group.

As expected from section 7.5.1 both architectures are heavily reliant on lexical features. Permuting the lexical features reduces the macro F1 prediction score by roughly 68% in both cases. In contrast, permuting the phonetic features impacts the macro F1 prediction score by roughly 2.5-4%, and permutation of the acoustic features impacts model performance less than half a percentage point at most. It seems clear that, with a few phonetic feature exceptions, the high performance of the STL and MTL models on region prediction seen in chapters 5 and 6 are predominantly driven by lexical features. Though the difference is small, it also appears that the STL model for region is a bit more reliant on phonetic features than the MTL model.

It's somewhat odd that phonetic features would have such a small impact on models trained to predict region. Regional phonetic differences are, after all, one of the primary pillars of traditional dialectology, and the basis for numerous dialectal divisions that sociolinguists and dialectologists often take for granted. There are a few possibilities for why phonetic features appear to contribute little to the overall predictive accuracy of the models as compared to lexical features. First, recall that each individual speech segment used as a data point in this dissertation is roughly 60 seconds long. It's possible that within this small time frame it is simply impossible to gather enough phonetic data to create an accurate representation of a speaker's phonetic landscape. This is likely particularly true for rarer phonemes– some of which are particularly important in drawing regional distinctions. Second, because the segments are so short for individual speakers, phonemes in this dissertation are not sub-divided in terms of surrounding phonetic context. Some of the regionally-specific phonetic phenomena identified in the dialectology literature are contextually dependent, and therefore not accounted for in the system deployed here. It's possible that if individual segments included larger chunks of speech, and if surrounding phonetic context for each phoneme were taken into account, phonetic features might be more useful to the models than they currently are.

Figure 7.9: Top 50 individual education features for STL architecture

7.6 Education

7.6.1 Individual feature importance

Figures 7.9 and 7.10 present a comparison of the top 50 most important features for the best performing STL and MTL models, respectively, trained to predict speaker education. Feature group is color-coded: lexical features are presented in grey, phonetic features in red, and acoustic features in turquoise.

As with most other traits examined in this chapter, the models trained to predict speaker education appear to rely predominantly on lexical features, with several acoustic and phonetic features mixed in to the top 50. The presence of acoustic features in the top 50 for both models seems on it's face somewhat odd, though recall from chapter

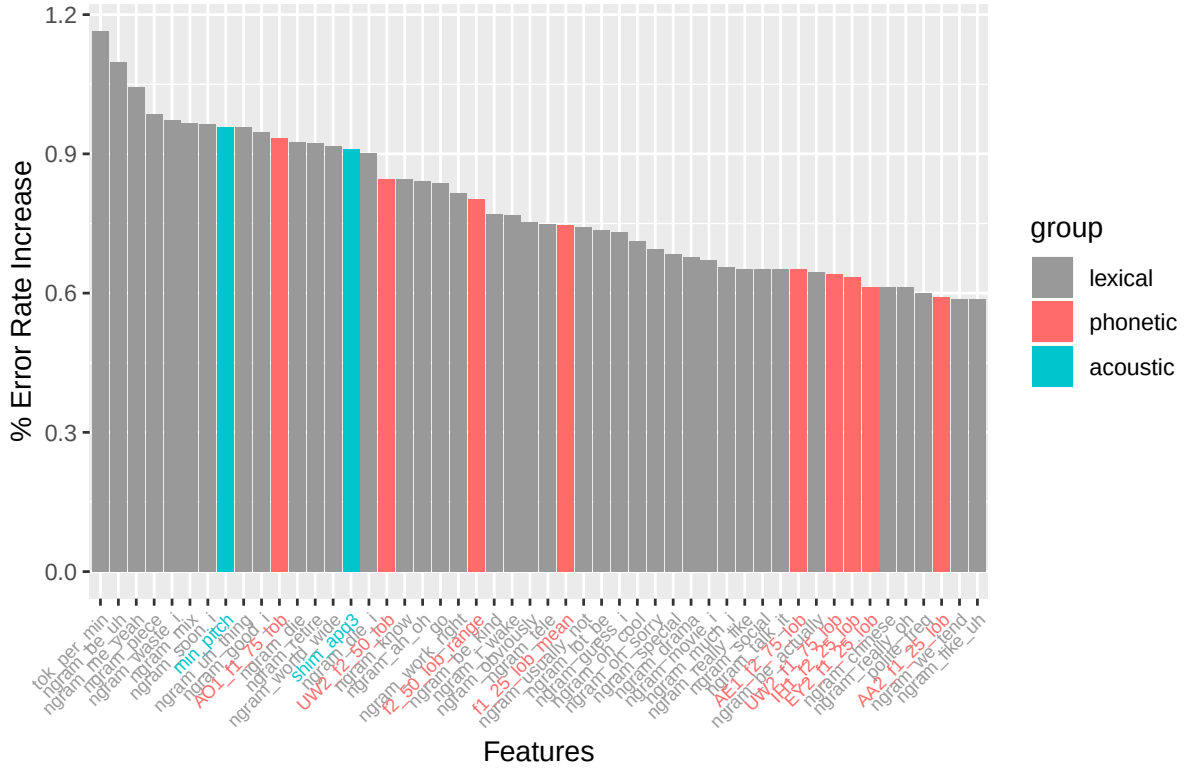

Figure 7.10: Top 50 individual education features for MTL architecture

4 that the education groups are somewhat unbalanced when it comes to speaker sex. The "no college" group, in addition to being by far the minority education group, is also predominantly male, whereas the "college" group is predominantly female and the "post-college" group is relatively balanced. It seems likely that the reliance on acoustic features in the models trained to predict speaker education is an artifact of this difference, in that male-indicative features may be useful in predicting the minority education class.

This sort of cross-over pattern in which models trained to predict trait A appear to take advantage of inter-group imbalance with respect to trait B by relying on input features directly related to trait B is found in the STL models for ethnicity and age as well. It is thus perhaps not surprising that a similar phenomenon would be observed here in the STL model for education. However, education is the only trait examined for which the corresponding MTL model also exhibits this pattern. For ethnicity and age,

the hypothesis was put forward above that while the MTL models represent relations between target and non-target traits via interactional features jointly developed in the shared hidden layer during training, the STL models need to do this more directly by relying on input features directly related to non-target traits, since they don't have any other way to represent the relation between target and non-target traits. If this hypothesis holds water, one would not expect the top 50 most important traits for the MTL model for education to include acoustic input features directly related to sex. That acoustic features do appear in the top 15 most important traits for the MTL education model may be an artifact of the difficulty of predicting speaker education. F1 macro scores for the STL and MTL models for education overall are around 0.75. These are some of the lowest overall scores found among the models trained in this dissertation. The closest overall scores are those for the age models, which also hover around 0.75. However, age is a 5-way discrimination task whereas education is a 3-way discrimination task. It may be that there is simply not a strong enough signal in the data as it relates to education for the sorts of jointly-developed, hidden-layer features hypothesized to be present in MTL models targeted at other traits to come to fruition in the MTL models targeted at education.

7.6.2 Feature type importance

Tables 7.9 and 7.10 present the overall permutation importance for each feature group calculated for Macro F1, as well as the mean of the individual ERI scores for each feature belonging to that group.

As with all other traits examined, with the exception of speaker sex, the models trained to predict speaker education are overwhelmingly reliant on lexical features to make accurate predictions. There does not appear to be a large difference in terms of the

Table 7.9: Feature group importance: Education STL

group	group error increase	mean individual error increase
lexical	58.05%	0.031%
acoustic	1.14%	0.447%
phonetic	-1.68%	0.187%

Table 7.10: Feature group importance: Education MTL

group	group error increase	mean individual error increase
lexical	55.33%	0.03%
acoustic	1.31%	0.1%
phonetic	-0.96%	0.023%

importance of each type of feature to the STL and MTL models, though the STL model appears slightly more reliant on lexical features than the MTL model. The phonetic feature group ERI for both models is negative, indicating that phonetic features had, at best, little to no impact on education predictions.

7.7 Discussion

7.7.1 Multi- vs. uni-factorial systems (research question 1)

The first research question laid out in chapter 1 was whether a multi-factorial system– that is, a system which incorporates features from multiple linguistic levels– meaningfully increases performance over a system including features of only one particular type. While no models were trained with solely one individual type of feature, the feature group permutation importance results from this chapter nonetheless allow us to speak to this question.

In no cases were the models tested completely unreliant on more than one of

the three feature type groups tested– that is to say, each model tested was somewhat reliant on at least two of the three feature types to some degree. Whether a model is *meaningfully* reliant on more than one type of linguistic feature however depends on one's definition of what is a meaningful reduction of error. For instance, those models that showed the strongest reliance on one individual feature group were the STL model for ethnicity (78.33%, 0.70%, and 0.25% model reliance on lexical, acoustic and phonetic features, respectively) and the MTL model for age (71.96%, 0.43% and -0.2% model reliance on lexical, phonetic, and acoustic features, respectively). If one determines a meaningful reduction of error rates to be a reduction of at least 1%, These two models may be considered to be wholly reliant on lexical features. Regardless, if one wants to achieve the highest classification performance possible, it does appear from the feature group permutation importance results that for all social traits examined one should include features from at least two of the three linguistic levels incorporated here.

In sum, if one desires the best possible results, a multi-factorial approach is best for all social traits. Depending on the trait in question and the tolerance for error in the application in question however, a multi-factorial approach may not always be worth the effort.

7.7.2 Feature type importance (research question 2)

The second research question laid out in chapter 1 addresses the relative predictive power of features from different linguistic levels on automated prediction of the five social traits examined here. Much of the existing computational literature on automated author profiling has focused on textual corpora rather than spoken corpora, and thus has focused almost entirely on the predictive power of lexical features for predicting author traits. However, most of the existing work on spoken corpora has tended to

focus on acoustic and to a lesser degree phonetic features, generally to the exclusion of lexical features. It is unclear from the existing literature whether the lexical features focused on in the textual corpora work would carry over to spoken corpora.

The results presented in this chapter provide quite clear evidence that this is indeed the case– lexical features appear to carry the majority of the predictive burden for all sociodemographic traits examined. However, the inclusion of phonetic, and to a certain degree acoustic features can, depending on the target trait, also provide a meaningful reduction in error rates. This is particularly true for ethnicity, region, and sex, for which the inclusion of phonetic features provided a decrease in error rate of between 2.5% - 5.5% depending on the model. There are some target traits however for which inclusion of phonetic or acoustic traits appears from the results presented above to be at best unnecessary and at worst detrimental. Education and Age in particular both benefited little from either acoustic or phonetic traits, and in the case of education the inclusion of phonetic features appears to have hurt overall performance.

Overall, the MTL and STL architectures appear to give largely the same weight to the different feature types. There does appear to be a slight tendency in the STL models to place a higher reliance than MTL models on input features that directly relate to non-target sociodemographic traits which may offer clues to target-trait classification, such as the relatively high importance placed on several acoustic features for the STL ethnicity model and the relatively high importance placed on certain toponyms for the STL age model.

7.7.3 Implications

The implications of the feature importance results presented above vary depending on the use case, data genre, and target sociodemographic traits to which an automated

speaker profiling system is applied. If it is crucial for an application to be as accurate as possible and the target trait is any of those examined in this dissertation other than education, it appears to be best to include all three types of features examined here– acoustic, phonetic, and lexical.

If however resources are scarce in terms of available development time, on-staff linguistic expertise, etc., the results presented above suggest that lexical features offer by far the highest cost-benefit 'bang for your buck.' Relying on lexical information alone in most instances will result in a system within a few percentage points of theoretical maximum predictive accuracy, the types of lexical features used above are relatively cheap and simple to extract with off-the-shelf, well-known programming tools, and extraction requires only a modicum of linguistic or software-specific know-how. Extraction of phonetic and acoustic features on the other hand require more involved linguistic expertise, and typically require the use of more specialized software tools such as Praat, vowel extraction and forced alignment software suites, and so on. In addition to these points, compiling and maintaining a corpus of auditory speech data is typically more work-intensive and takes up more storage space than compiling and maintaining a corpus of textual data, and any pre-existing corpora of language data that most non-academic institutional entities may want to take advantage of for training purposes are more likely to be in textual rather than auditory format.

One caveat to relying solely on lexical features however is the potential pitfall of covert over-fitting with respect to lexical items in a specific corpus. Lexical features in general are likely to be more vulnerable to domain change, topic variation, and so on than acoustic or phonetic features. In order to deploy a real-world system heavily reliant on lexical features which is broadly applicable, one should attempt to mitigate the possibility of covert over-fitting by using training data from a corpus that is better balanced across topics and text types than that used in this dissertation.

Chapter 8: Discussion

Having now presented the results of both the MTL and STL experiments and examined the relative importance of feature-type for each architecture, we can begin to draw some conclusions about the relative benefits and detriments of MTL vs. STL architecture.

8.1 Overall Performance

The first thing to note is that the performance of both the STL and the MTL models presented in this dissertation is quite good in comparison to similar existing work. The closest comparison to the experiments performed in this dissertation that I'm aware of in the existing automated speaker profiling literature is Gillick (2010). Using a similar subset of the same corpus relied upon in this dissertation– the 2008 NIST Speaker Recognition Evaluation data set– Gillick trained Margin Infused Relaxed Algorithm (MIRA) classifiers to predict sociodemographic speaker information along the same five demographic axes focused on in this dissertation: sex, ethnicity, age, region, and education. In all cases other than age, the classification schemas used by Gillick were identical to those used in this dissertation. For age, Gillick used a four-way bucketing schema (20-29, 30-39, 40-49, 50+) whereas this dissertation used a slightly more fine-grained five-way bucketing schema (18-25, 26-35, 36-45, 46-55, 56+). For each sociodemographic trait, Gillick used as features a binary representation (presence/absence) of the top 2000 bigrams according to information gain. This is very similar to the lexical

features used in this dissertation, though rather than solely bigrams, the lexical features in this dissertation are a selection of the top 2000 ngrams from the set of both bigrams and unigrams.

The main differences between the experiments performed in Gillick (2010) and those performed here are as follows:

- Gillick (2010) relies solely on lexical features, whereas the models presented in chapters 5 and 6 rely on lexical, phonetic, and acoustic features.
- Gillick (2010) employs MIRA classifiers to perform sociodemographic trait prediction, whereas this dissertation uses various forms of Multi-Layer Perceptron Neural Networks.
- Gillick (2010) does not report any sort of feature preprocessing, whereas this dissertation applies various preprocessing steps to the raw features such as reduction of highly correlated feature clusters, Yeo-Johnson power transformation, etc. as described in chapter 5.

Table 8.1 compares the accuracy rates for the MIRA classifier for each trait reported by Gillick (2010) to accuracy rates[1] for the STL and MTL models presented in chapters 5 and 6 of this dissertation.

Table 8.1: Predictive accuracy for Gillick (2010) and the STL and MTL models

Trait	Gillick 2010	STL	MTL
Sex	82%	97.3%	98.1%
Ethnicity	72%	87.3%	87.8%
Age	65%	75.9%	76.9%
Region	60%	83.0%	82.9%
Education	67%	79.4%	80.2%

In all cases, the STL and MLP models reported in this dissertation perform between 10-20% better than the models reported by Gillick (2010). The lowest disparity between the models presented here and those presented by Gillick is for age, but again

[1]Accuracy is used for comparison here rather than a form of F1 as this is the only evaluation metric reported by Gillick (2010).

recall that the classification schema used in this dissertation treats age as a five-way classification problem whereas Gillick treat age as a four-way classification problem.

The improved classification accuracy for the STL and MTL MLP models over Gillick's MIRA models may be due to a number of factors. Perhaps the most obvious is the additional phonetic and acoustic features included in the STL and MTL MLP models. The inclusion of acoustic and phonetic features likely is responsible for some portion of the increased predictive accuracy of the MLP models– particularly in the case of sex– however the low importance of phonetic and acoustic features for prediction of most of the sociodemographic traits examined here make it unlikely that these additional features are the main driving force behind the disparity. Another possible factor is the implementation of the lexical features included in the models. The STL and MTL MLP models presented in this dissertation use a combination of unigram and bigram features rather than solely unigram or solely bigram features precisely because preliminary experimentation suggested that such a combination provided slightly better results (on the order of 1-5 percentage points) than using purely unigrams or purely bigrams. However, as with the additional acoustic and phonetic features, this seems unlikely to be the main factor responsible for the disparity. It seems likely that the driving force behind the substantial gains in predictive accuracy of the MLP models over the MIRA models is predominantly a function of model type. Though admittedly more computationally intensive, Multi-Layer Perceptron models may simply be better suited to this type of task than MIRA models.

8.2 Multi-Task vs. Single-Task Learning

As described in chapter 6, evaluation metrics for the STL and MTL models are quite close for all sociodemographic traits examined, but the MTL models generally edge out

the STL models for most evaluation metrics for most traits by about a percentage point or so. There appears to be a slight tendency for MTL models to outperform STL models particularly in identification of well-represented (majority) classes, though there also appears to be a very slight tendency to under-perform STL models on identification of the most extreme under-represented classes.

Despite the MTL models generally outperforming the STL models, the difference in terms of evaluation metrics between these two types of architectures was not nearly as large as expected. A host of research detailed in chapter 2 suggests that MTL frameworks tend to out-perform STL frameworks somewhat substantially when applied to highly related tasks, yet this does not appear to be the case for the models reported in this dissertation. It's important to dig into why this might have been.

One clue as to the underlying factor for the relative similarity between the STL and MTL models reported here may lie in the feature importance results detailed in chapter 7. It became clear in chapter 7 that, though a few individual phonetic or acoustic traits might be highly important to prediction depending on the sociodemographic trait of focus, lexical features as a group were by far the most important feature block, contributing as much as 78% of the error reduction rates depending on the trait of focus. Recall also that the lexical features used in models targeted at predicting each individual trait were engineered so as to have specific importance to that particular trait. That is to say, the lexical features used in the ethnicity predicting models were selected specifically so as to be useful in distinguishing between ethnicities, the lexical features used in the age-predicting models were selected so as to have maximal relevance to distinguishing between age categories, and so on. The choice was made to include only ngram features that were relevant to the target trait because results from initial experiments including all lexical features for all traits were particularly poor and quite close to naive baseline levels– likely due to the high dimensionality of

the resulting feature vector.[2] For the MTL models, that meant that, though those models were attempting to learn representations for sex, age, ethnicity, region, and education simultaneously, they only received lexical features that were guaranteed to be relevant to one of those five traits during training. In light of the importance of lexical features to sociodemographic trait prediction, it is reasonable to hypothesize that, lacking lexical features relevant for predicting the four non-target traits, those output layers that were dedicated to predicting the non-target traits were unable to establish a strong enough representation of their particular assigned sociodemographic trait so as to impact model performance any more than a moderate few tenths of a percent. In other words, the network-regularization benefits of the MTL architecture may have been mostly nullified by the failure to deliver relevant lexical feature information to the non-target sociodemographic heads of the network.

In light of this possibility, two additional experiments were performed that extended the MTL design detailed in chapter 6 in order to address this potential issue. Both experiments modified the network design in order to provide those sections of the MTL network dealing specifically with non-target sociodemographic traits with lexical information relevant to those traits.

8.3 Multi-Task Learning Extension Experiments

8.3.1 Dense embeddings as lexical features

In this experiment, the trait-specific binary ngram features used in the initial MTL experiment were replaced with trait-agnostic, segment-specific dense embeddings. Two

[2]After eliminating duplicates and removing highly correlated clusters, including the top ngrams for each of the 5 traits results in a feature vector with roughly 7,000 dimensions. Including only those ngrams relevant to the target trait however results in a feature vector of roughly 2,200 dimensions.

different methods for obtaining dense embeddings from each speech segment were explored: 1) using a version of Google's pre-trained Universal Sentence Encoder (Cer et al., 2018), and 2) using a standard Doc2Vec model trained on the corpus as a whole.

The idea behind this experiment was that, by delivering a dense embedding representation of the words used in a given speech segment, one could avoid the problem of high dimensionality in the feature vector while still providing a reasonably full picture of the lexical items used in a speech segment to the network. The sections of the MTL network specifically focused on individual sociodemographic trait prediction could then learn to attend to those dimensions of the embedding vector that were relevant to those specific traits, thereby gaining the relevant, trait-specific lexical information that had been previously only available for the target sociodemographic trait.

Table 8.2 presents the macro F1 scores for the MTL models trained to predict each sociodemographic trait of focus using both Universal Sentence Encoder-derived embeddings and Doc2Vec-derived embeddings and compares them to the macro F1 scores for the initial binary ngram MTL models trained to predict those same traits as reported in chapter 6. Hyper parameters were re-tuned for each of the new approaches using the same methodology as was used for the original binary ngram MTL models prior to final model training. As with the ngram MTL models, results presented in table 8.2 below represent the average macro F1 score over five separate training runs.

Table 8.2: Evaluaton metrics for ngram, Doc2Vec, and USE MTL models

trait	Ngram_F1	Doc2Vec_F1	USE_F1
sex	0.980	0.951	0.959
eth	0.785	0.419	0.403
age	0.758	0.345	0.367
reg	0.818	0.363	0.412
edu	0.758	0.420	0.355

In all cases, the F1 scores for models using dense embeddings were worse than those for models using specifically selected ngram features. For all traits except sex, the embedding models were much, much worse. The relatively minimal impact of embedding vs. ngram features for sex is likely related to the lower reliance on lexical features for models predicting sex in general.

It appears that for MTL models performing this type of task, it is indeed best to rely on ngram features specifically selected to be relevant to the target trait, despite these features potentially having little relevance to the non-target traits. The poor performance of the embedding models may stem from the nature of the embeddings themselves. Dense embeddings necessarily reduce the distinction between semantically similar lexical items, and in some cases subtle differences between what may be otherwise semantically similar items can be crucial to distinguishing between sociodemographic categories. For instance, the lexical items "husband" and "wife" are semantically identical in most respects other than gender, and the use of one or the other is unlikely to make much difference to a dense embedding representing the speech segment in which it occurs. However, when looking at usage patterns of these two lexemes in the corpus, it is a near certainty that if a speaker uses the word "husband" that speaker is a female, and vice versa. Likewise the terms "husband" and "boyfriend" are semantically similar in most respects, yet use of the former is far more likely to indicate a speaker from one of the older age groups than one of the younger age groups. As a final example, dense embeddings are unlikely to distinguish much between individual toponyms, yet mentions of "bay area," for instance, typically indicate a speaker from the Western region, toponyms belonging to the northeast typically denote northeastern speakers, and so on.

8.3.2 Trait-specific lexical feature delivery

The second of the MTL extension experiments attempted to address the lack of relevant lexical information available to the non-target sections of the MTL models by moving delivery of the lexical features out of the shared layer and into the trait-specific layers. Acoustic and phonetic features were delivered to the shared hidden layer as before. For each trait, the output of the shared hidden layer was then concatenated with a vector representing the presence or absence of the ngrams relevant to that particular trait, and then passed on to the hidden layer that dealt specifically with that particular trait. In other words, all sections of the MTL network received the same phonetic and acoustic information, but each trait-specific section of the network only received lexical information relevant to that particular trait. Tables 8.3 through 8.7 compare the evaluation metrics of the initial "vanilla" ngram MTL models with those using this trait-specific ngram delivery design for each trait.

Table 8.3: Comparison of vanilla and trait-specific MTL models for sex

condition	f1_macro	f1_weighted	acc
vanilla	0.980	0.981	0.981
t-spec	0.966	0.967	0.967

Table 8.4: Comparison of vanilla and trait-specific MTL models for ethnicity

condition	f1_macro	f1_weighted	acc
vanilla	0.785	0.884	0.878
t-spec	0.821	0.893	0.891

Table 8.5: Comparison of vanilla and trait-specific MTL models for age

condition	f1_macro	f1_weighted	acc
vanilla	0.758	0.771	0.769
t-spec	0.750	0.764	0.763

Table 8.6: Comparison of vanilla and trait-specific MTL models for region

condition	f1_macro	f1_weighted	acc
vanilla	0.818	0.830	0.829
t-spec	0.819	0.825	0.825

Table 8.7: Comparison of vanilla and trait-specific MTL models for education

condition	f1_macro	f1_weighted	acc
vanilla	0.758	0.800	0.802
t-spec	0.841	0.861	0.862

For sex, age, and region, the trait-specific ngram delivery design either had minimal impact or performed worse than the vanilla ngram delivery design. For ethnicity and education however, macro F1 scores improved somewhat dramatically. In the case of education, macro F1 score improved by over 8 points, along with a corresponding 6 point improvement for weighted F1 and and overall accuracy. That the gains for these two traits are larger in macro F1 than the other evaluation metrics suggests that the trait-specific ngram delivery design was of particular help in identifying speakers belonging to minority classes. Figures 8.1 and 8.2 present normalized confusion matrices averaged over the five training runs for the trait-specific ngram delivery models for ethnicity and education, respectively.

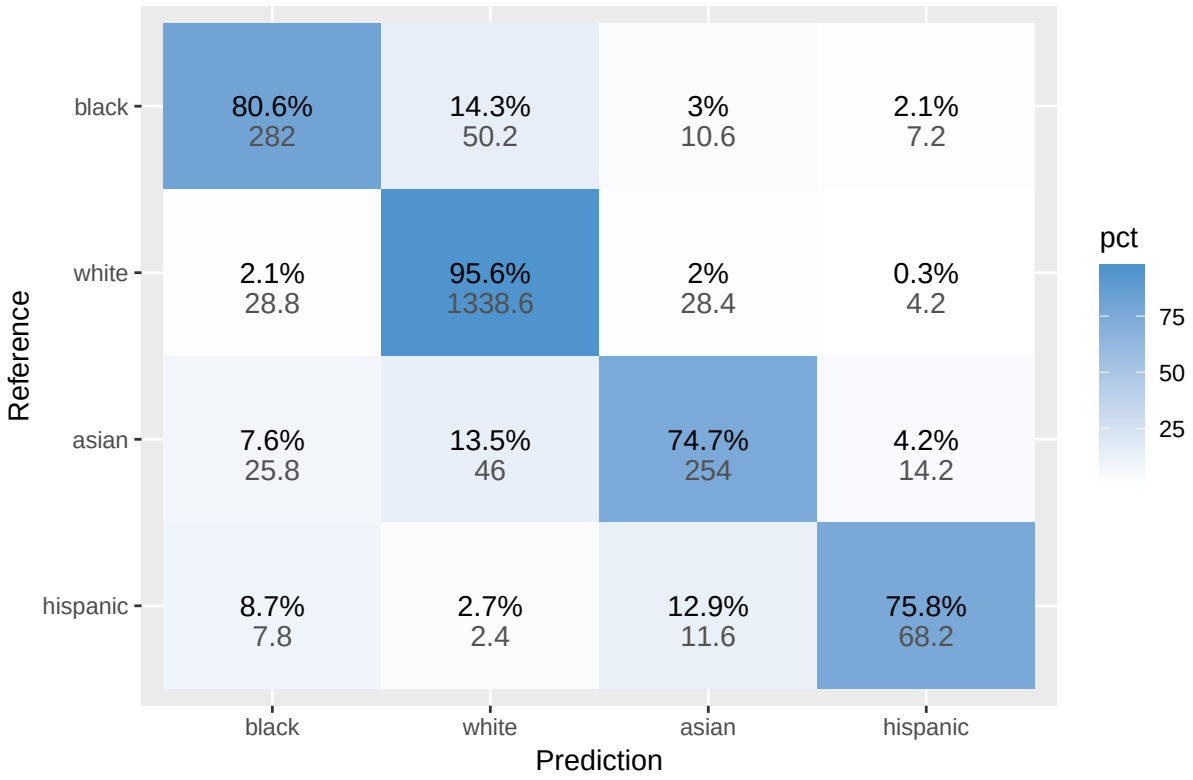

Figure 8.1: Confusion matrix for trait-specific MTL Ethnicity models

The confusion matrices in figures 8.1 and 8.2 confirm that the major areas of improvement for the trait-specific ngram delivery design for these two features were indeed in identification of minority class speakers, at the slight expense of predictive accuracy for the majority class speakers. For ethnicity, the major difference is the identification accuracy of Hispanic speakers– jumping from 49.7% accuracy in the vanilla MTL models to over 75% accuracy in the trait-specific ngram delivery models. Likewise identification accuracy for "no college" and "post-college" speakers in the education models improved from 68.5% and 72.3% to 80.9% and 81.7%, respectively. An examination of the confusion matrices for the other three traits reveals a similar pattern wherein the most under-represented classes for the respective sociodemographic trait generally experience a slight boost in identification accuracy and the most over-represented classes generally experience a slight decline in identification accuracy. Though not reported here for reasons of space, examination of the non-target output layers of the MTL models also

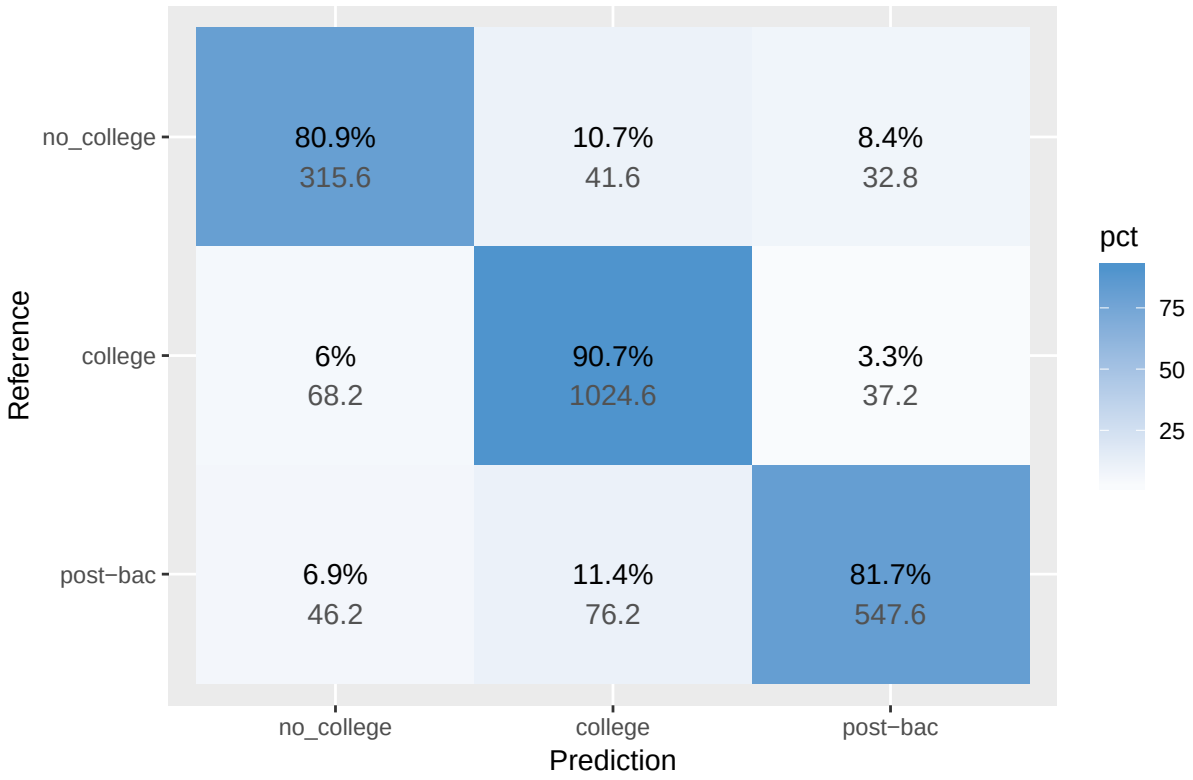

Figure 8.2: Confusion matrix for trait-specific MTL Education models

show a (predictably) substantial increase in the predictive accuracy of the non-target sociodemographic trait prediction for the trait-specific ngram delivery architecture across the board. These pieces of evidence support the notion that the trait-specific ngram delivery design increases the efficiency of the non-target MTL heads, thereby increasing the effectiveness of the regularization effects of the MTL architecture on the model as a whole and boosting performance in those situations where network regularization would have the strongest effect– namely in curbing over-prediction of the majority class and boosting identification of those sociodemographic classes which are the most severely under-represented in the data and which present the most difficulty to vanilla STL and MTL models.

8.4 Conclusions (Research Question 3)

Given the findings in chapters 5, 6, and 7 and the discussion presented above in this chapter, there are several main conclusions regarding the relative performance of STL and MTL models that are worth outlining here in detail. These conclusions directly address research question 3 as laid out in chapter 1: Can a multi-task learning approach provide significant gains in accuracy over a system in which each speaker trait is predicted in isolation?

First, it is apparent from the comparison of the STL and vanilla MTL performance metrics that the MTL architecture does indeed generally provide better performance on sociodemographic trait prediction than the STL architecture, regardless of one's preferred evaluation metric. However, the increase in performance for the MTL models for most traits is minimal (and in the case of region, nonexistent), and may not warrant the increased complexity of the MTL architecture.

Second, though the inclusion of phonetic and acoustic features do provide an increase in performance for most sociodemographic traits (particularly in predicting speaker sex), lexical features as a group are by far the most important for obtaining accurate results. In addition to their overwhelming importance in all of the models addressed in this dissertation, lexical features are generally cheaper and easier to extract than phonetic and acoustic features, and require less specific linguistic knowledge and tools to do so.

Third, in terms of representing the lexical content of a speech segment, binary (presence/absence) ngram features specifically selected to have relevance to the target trait provide substantially better results than trait-agnostic embeddings. As discussed above, this is likely due to the fact that embeddings by design reduce the distinction

between semantically similar lexical items that, despite their overall semantic similarity, may provide crucial clues to sociodemographic class identification.

Fourth, in terms of feature delivery, delivering all binary ngram features that have relevance to all traits predicted in the output layers of an MTL model together in the initial shared layer of an MTL model results in prediction performance close to naive baseline. This is likely due to the high dimensionality of the resulting feature vector. Presenting just those ngram features which have relevance to the target trait in the initial shared layer of an MTL model results in overall excellent prediction performance, edging out STL models using the same features and training data, and outstripping the most similar work in the automated speaker profiling literature (Gillick, 2010) by roughly 16 percentage points in terms of overall predictive accuracy on average. Moving delivery of ngram features from the shared section of the network to the trait-specific sections of the network and providing each output layer with ngram features designed for their respective sociodemographic traits results in even higher overall accuracy for some sociodemographic traits (particularly for education, which receives a 6% increase in overall accuracy), and generally boosts model performance in terms of identification accuracy of minority classes.

The above conclusions lead to the following actionable recommendations for those considering the use of Multi-Task Neural Networks in predicting sociodemographic speaker traits.

First, if feasible, one should attempt to include features from the phonetic, acoustic, and lexical realms in systems designed to predict sociodemographic speaker traits. If however working on a project with time/resource constraints, one should focus on the extraction of target-trait-relevant binary ngram features, as these provide the highest cost/benefit ratio.

Second, MTL models are trickier to design, have more hyper-parameters to optimize, take longer to train, and are more computationally intensive than their STL counterparts. If the goal of a project is to obtain the absolute best predictive performance possible, an MTL framework should be used over an STL framework. However, If a percentage point here or there isn't crucial and/or resources are limited, STL frameworks are probably adequate for most tasks. The exception to this is in the prediction of education. The trait-specific ngram delivery flavor of the MTL architecture resulted in a net 6.8% increase in predictive accuracy over the corresponding STL model. Such an increase may warrant the increased complexity, training time, etc. that is inherent in the MTL architecture.

Chapter 9: Contributions, Limitations, and Future Work

This chapter provides an overview of the contributions this dissertation has made to the field of automated speaker profiling as well as some limitations of the present work and potential future avenues of exploration.

9.1 Contributions

As far as I'm aware, this dissertation represents the first attempt to apply multi-task learning to automated speaker profiling tasks. Given the wealth of evidence from other fields for the power of multi-task learning on related classification tasks, this exploration has been long overdue. A summary of the classification accuracy and macro F1 scores for all models trained on all tasks is presented in tables 9.1 and 9.2, respectively. The best performance scores for each trait are bolded.

Table 9.1: Summary of classification accuracy for all STL and MTL models

model type	sex	eth	age	reg	edu
STL	0.973	0.873	0.759	**0.83**	0.794
MTL (vanilla)	**0.981**	0.878	**0.769**	0.829	0.802
MTL (Doc2Vec)	0.952	0.632	0.409	0.422	0.518
MTL (USE)	0.959	0.649	0.403	0.458	0.455
MTL (t-spec)	0.967	**0.891**	0.763	0.825	**0.862**

Table 9.2: Summary of macro F1 scores for all STL and MTL models

model type	sex	eth	age	reg	edu
STL	0.973	0.79	0.743	**0.82**	0.752
MTL (vanilla)	**0.98**	0.785	**0.758**	0.818	0.758
MTL (Doc2Vec)	0.951	0.419	0.345	0.363	0.42
MTL (USE)	0.959	0.403	0.367	0.412	0.355
MTL (t-spec)	0.966	**0.821**	0.75	0.819	**0.841**

The results presented in chapters 5 through 8 and summarized in tables 9.1 and 9.2 demonstrate that multi-task models outperform single task models regardless of evaluation metric on speaker classification along four of the five social traits examined (sex, ethnicity, age, education). This finding leads to the actionable recommendation detailed in chapter 8 that those wishing to deploy automated speaker profiling systems in real world contexts where classification accuracy is paramount would be well served by adopting the type of multi-task model design detailed here. However, the comparison between the multi-task models and single-task models also makes clear that while multi-task models consistently outperform single task models, the improvement in classification accuracy for most traits is somewhat minimal from a practical standpoint. The best performing MTL models outperformed the STL models in classification accuracy by roughly 2% on average. Though this improved performance of MTL over the STL model design is important from a theoretical perspective, the magnitude of said improvement may not be worth the added design complexity depending on one's use case. As discussed in chapter 8, single-task systems should be sufficient for those deploying automated speaker profiling systems in contexts where time, resources, and/or expertise are limited.

Beyond the comparison of single-task and multi-task frameworks, a further contribution of this dissertation to the field of ASP is that, so far as I'm aware, the classifica-

tion accuracy of the best performing MTL models presented in chapter 8 represent the highest performance on speaker education and speaker ethnicity prediction that have been achieved on conversational speech to date. Ethnicity and education are particularly under-examined traits within the field of automated speaker profiling, perhaps due to task difficulty, yet this dissertation has demonstrated that the type of MTL models detailed here are capable of classifying speakers with regard to these traits with accuracy rates in the mid to high 80% range (86.2% for education, 89.1% for ethnicity). The closest performing models I'm aware of detailed in the ASP literature come from Gillick (2010), with accuracy rates of 67% for education and 72% for ethnicity. This represents a major improvement in performance on these tasks, and should serve as a jumping off point for future ASP work concerned with these traits.

Performance metrics of the MTL models on speaker sex classification are also some of the best figures reported for spoken data, within 0.5% of the maximum classification accuracy reported in the ASP literature. Recall that the highest performance on sex classification using spoken data to date come from Hu et al. (2012), who reported 98.65% accuracy rates in a binary sex prediction task. However, the data-set used in this study consisted of extremely high quality laboratory recordings of speakers producing 77 digit sequences. The MTL models detailed in chapter 6 achieved an accuracy rate of 98.1% on conversational telephone data– a rather remarkable result given the medium. As far as I'm aware, this represents the highest performance achieved thus far on speaker sex classification using conversational, non-laboratory speech data.

In addition to the extremely high performance of the MTL models presented here, one of the prime contributions of this dissertation to the field of ASP are the findings related to the importance of different feature types. As discussed in chapters 1 and 2, the majority of automated systems attempting to profile speakers using spoken data tend to focus on phonetic and acoustic features. The analysis presented in chapter 7

however clearly shows that lexical features are the driving force behind the high performance of the models reported here, responsible for up to 78% of the reduction in error rates. While the high level of model reliance on lexical features relative to phonetic and acoustic features may be somewhat surprising to sociophoneticians and those working in traditional automated speaker profiling, a great deal of computational corpus work over the years has demonstrated the power of lexical predictors for distinguishing between various social trait categories. Boulis and Ostendorf (2005) for example achieved a classification accuracy rate of 92% on a binary sex classification task performed on transcripts of conversational telephone speech data using presence/absence lexical features similar to the informative ngram features used here. Rao et al. (2010) demonstrated that Twitter users may be classified with a high level of accuracy according to gender (male/female, 72% accuracy), age (above/below 30, 74% accuracy), regional origin (north/south India, 77% accuracy), and political orientation (Democrat/Republican, 83% accuracy) using a combination of lexical and orthographic features. Nguyen et al. (2013) have even demonstrated that reasonable accuracy (micro F1 scores between 0.85-0.87) can be achieved in classifying Twitter users by age-group and life-stage based on unigram lexical information alone. It is safe to say that the power of lexical features on speaker/author classification tasks is well known in the corpus-based computational literature. The present work demonstrates that this finding carries over into automated speaker profiling, and should motivate researchers working in the field of automated speaker profiling moving forward to broaden the scope of higher-level features considered and increase the level of attention paid to potential lexical predictors, regardless of the social trait(s) of focus.

Beyond model design and implementation ramifications, this dissertation also implicitly presents an interesting methodological point that future research in ASP should take note of. One of the major difficulties in performing speaker classification tasks using spoken data is the relative paucity of (conversational) spoken data corpora available

for model training. This fact is likely primarily responsible for the relatively scant automated profiling work done on spoken as compared to textual data, and may also be responsible for the general trend in automated speaker profiling work (as compared to automated author profiling work) to use model types that require fewer data points for effective training than the types of neural network models used here. This dissertation addresses this point by performing a randomized speech segment chunking procedure that artificially boosts the number of data points ten-fold. The high performance of the models presented in preceding chapters speaks to the fact that this chunking procedure had little to no detrimental effects on speaker classification, and can be an effective strategy for increasing the number of data points available for model training.

Relatedly, the performance metrics of the models presented here speak to the fact that the 60 second speech segments extracted from the original five-minute conversational recordings contain sufficient linguistic information to achieve high classification accuracy in prediction tasks examining all five of the social traits investigated in the preceding chapters. Though longer speech segments (and thus more linguistic information) would likely increase performance somewhat, it appears that 60 seconds of conversational speech is sufficient to achieve quite high classification accuracy on the profiling tasks undertaken in this dissertation.

9.2 Limitations

Prior to addressing the potential future avenues of exploration that this dissertation points to, it is important to point out several limitations that may have impacted the present work.

First, it's important to note that though the models presented here achieve particularly good results on the NIST SRE data, the generalizability of these models to

other corpora and other genres is unknown at this time. As pointed out in chapter 7, the high importance some of the models place on individual content lexemes which are seemingly unrelated to the task at hand may be an indicator that these models have covertly over-fit the corpus used for training and testing (that is to say, though the models do not appear to unduly rely on idiosyncrasies of the training versus the test sets, they may be reliant on certain idiosyncrasies present in the NIST SRE data set as a whole). If this is indeed the case, this coupled with the high model reliance on lexical features makes it questionable that the models presented here would generalize well to conversational corpora from different time periods, non-conversational spoken corpora, or conversational corpora including different conversational topics than those present in the NIST SRE data-set. That said, out-of-domain performance degradation is typical in any sort of machine learning situation, and the network design presented in the preceding chapters does include regularization strategies such as drop-out and batch-norm (as well as the added regularization effects of the MTL design itself) to mitigate this somewhat. Furthermore, while the models themselves may not generalize well, I see no reason why the model design, feature engineering, and training procedures used here would not generalize well to other corpora. In other words, while it is questionable that the specific models trained here would perform as well on test data from other corpora, models trained on any particular conversational corpus using the same methodology as laid out in this dissertation would likely perform similarly well on test data from that same corpus.

It should also be noted that the relatively low importance assigned to phonetic and acoustic feature types by the models presented here may simply be a result of not including important phonetic and acoustic features. Several features potentially important to speaker classification were discussed in chapter 2 that were not included as features in the models presented above due to time constraints. It is likely that inclusion of such features (e.g. contextually specific representations of certain vowels

such as pre-nasal /æ/, information related to consonants such as voice onset time of certain stop phonemes, specific alternations mentioned in the sociolinguistic literature such as IN/ING variation, and so on) would boost both overall performance as well as acoustic/phonetic feature type importance.

A further limitation of the present work stems from the nature of the feed-forward neural network models used. Such models necessitate treating each speech segment as a "bag of features," and thus necessitate operationalizing most features as an average taken over the entire duration of the speech segment. As such, these models are not able to take into account any sort of potentially useful temporally or contextually dependent phenomena. If for instance a particular pronunciation of a particular word is diagnostic of a certain demographic category (yet this particular pronunciation is not indicative of a systematic peculiarity of the phonetic system of these individuals at large), this is information unable to be captured in the current methodology since all phonetic, acoustic, and lexical features extracted are the result of averaging all examples encountered throughout the course of the speech segment. In short, the current methodology does not allow for providing the models with feature conjunction information of the type "acoustic/phonetic phenomenon X was encountered during productions of lexeme Y throughout this speech segment."[1]

Finally, it may be important to consider the fact that the type of permutation feature importance testing performed in chapter 7 does not necessarily convey the same information that would be gained via ablation testing. As such, while we can measure reductions in error rates depending on the particular features and feature groups that are permuted, we can't necessarily draw conclusions regarding how a given

[1]For example, producing a place name, food term, etc. associated with Hispanic culture in such a way as to be consistent with the phonetic inventory of Spanish rather than the typical Americanized pronunciation is likely highly indicative of Hispanic ethnicity. Encountering such a pronunciation would likely be a "knock-out" feature for most human listeners, leading them to conclude that the unknown speaker identifies as Hispanic, whereas a system examining the average phonetic landscape over the entire speech segment would likely miss such a momentary shift.

model would perform in the absolute absence of these particular features/feature groups. Relatedly, it's possible that the high importance placed on the lexical feature group for all models examined was driven not necessarily by a high degree of information conveyed by these features, but rather by the high degree of noise conveyed by permutation of these features. In other words, it's possible that the extreme performance degradation observed for models receiving randomly permuted lexical features may not be due to a loss of informational signal, but rather due to the introduction of approximately 2,000 features containing nothing but random noise. It's therefore possible that a model trained and tested on solely phonetic and acoustic information would perform better than a model trained on phonetic, acoustic, and lexical information and tested with phonetic, acoustic, and permuted lexical information. Beyond these points, the nature of individual feature permutation testing means that clusters of important yet somewhat mutually redundant features will be underestimated in terms of their individual feature importance, as discussed in chapter 7.

9.3 Future Work

This dissertation points to a number of interesting avenues for future exploration.

First, though this dissertation employed speech segments of roughly 60 seconds as the atomic unit of analysis, this chunk length was somewhat arbitrarily chosen. It would be an interesting and worthwhile endeavor to experiment with chunk length to divine the smallest length of speech segment necessary to perform these profiling tasks without substantial performance degradation. The smaller a chunk one needs to rely on for these sorts of tasks, the more data points one can mine from corpora containing longer conversational segments and the more use cases such a system could be applied to. It may also be beneficial to explore the use of chunk size as a hyper-parameter

co-dependent on the corpus used and task priorities in order to find the optimal chunk size for a given data-set and classification problem.

Continuing in the methodological vein, though this dissertation trained neural networks using a feed-forward network design, it would be interesting to extend the multi-task methodology presented here for use with recurrent neural network designs. Such recurrent networks would not necessarily need to rely on specific speech segment chunk lengths, which could potentially widen the range of use cases to which these models may be applied. Such networks would also be able to take into account information related to the sequential order of features encountered within a given speech segment rather than treating each speech segment as essentially a bag-of-features.

One particularly important and potentially useful avenue of future research would be to explore broader, more general methods of representing lexical information to the predictive models than the informative ngram approach used here. Informative ngram features were found to be highly predictive of social category distinctions both here and in Gillick (2010), yet explicitly selecting particular ngrams relevant to particular social traits is somewhat inelegant and must be done anew for each novel trait that this sort of system is applied to. The dense embedding extension experiments discussed in chapter 8 were designed specifically to provide a broader and more trait-agnostic representation of lexical features to the MTL models, though results for these MTL models were quite a bit lower than for the MTL models using informative ngram features. One possibility not explored here is representing the lexical choices made by speakers via character-based language models. Such an approach would have the added benefit of directly modeling morphological phenomena, which though potentially predictive was not examined in any way in the current work. The ability to represent morphological as well as lexical information within the predictive models may be of particular use in predicting traits such as level of education.

Extensibility of the current methodology to multi-lingual settings could also be an interesting direction of future research. The acoustic features included in the methodology presented here should be largely language-agnostic. Phonetic and lexical features would of course need to be tailored to the phonetic and lexical inventory of whichever new language(s) this methodology would be extended to work with, yet I see no particular reason why the basic underlying concepts of measuring vowel point representations, vowel space diagnostics, and extracting representations of informative ngrams would not be of use in non-English ASP settings.

One relatively minor methodological point that may bear further examination is the manner in which phonetic representations are obtained. As noted in chapter 3, this dissertation employed a forced aligner to automatically segment the sound files at both the lexical and phonetic levels. Forced aligners such as that used here are reliant on both a transcript and an underlying pronunciation lexicon, both of which may be prone to errors. Erroneous phonetic transcriptions resulting from incomplete pronunciation profiles in the lexicon and inaccuracies in the transcript may be reduced by relying on a phonetic recognition system[2] to extract phonetic representations, thus making such representations lexicon and transcript agnostic.

Something else to look into is whether or not there is a difference in generalizability of the STL vs. the MTL models. The added regularization effects and auxiliary shared features of the MTL network design could very well make MTL speaker profiling systems more robust to domain and genre changes than systems using an STL design, which would be another point in favor of MTL in situations where broad applicability of speaker profiling systems is desired. This is an open question, and one that is ripe for future investigation.

[2]See e.g. Alsharhan and Ramsay (2019) for a recent example of such a continuous phonetic recognition/transcription tool.

Another potentially worthwhile experiment would be to examine the robustness of the current methodology to malicious signal manipulation aimed at deception. The high reliance on lexical features[3] of the models presented above would seem to make it likely that such models would be relatively robust to the types of acoustic and phonetic manipulation commonly employed for disguising one's speech, which could be potentially useful in law enforcement contexts (e.g. narrowing down the pool of suspects for an individual making a bomb threat).

One particularly interesting area for exploration may be the disconnect between what the models presented here find important (i.e. the high importance placed on lexical features) and what human listeners tend to find important when asked to classify unknown individuals along sociodemographic axes using spoken data. Purnell et al. (1999) for instance find that human listeners are able to correctly identify speaker ethnicity solely based on hearing a recording of a single word ("hello") more than 70% of the time. Clearly in this case listeners are not attending to lexical cues, but rather phonetic and/or acoustic cues. In comparison, the vanilla MTL models trained to predict ethnicity dropped from roughly 88% classification accuracy to roughly 30% classification accuracy when presented with permuted lexical data (i.e. presented with data containing reliable information solely for the phonetic and acoustic features). It therefore seems likely that human listeners are able to pick up subtle phonetic, acoustic, and potentially paralinguistic cues that are not included in the feature sets used here. Again, this suggests that it may be fruitful to explore the types of features which are known within the sociolinguistic literature to vary in socially meaningful ways yet which were not included in the present analysis. Of particular interest would be an

[3]While speakers deliberately attempting to disguise their speech may add, avoid, or vary certain high-saliency lexical items associated with particular social groups (e.g. the northern California intensifier "hella", slang words stereotypically associated with AAE or CE, etc.), most of the lexical cues on which these models rely are likely below the level of consciousness and not subject to intentional manipulation, such as the frequency and distribution of certain discourse markers, pause fillers, pronoun distributions, etc.

examination of certain paralinguistic cues related to prosody and stress patterning which are known to pattern socially,[4] as I am unaware of any work in automated speaker profiling so far which takes such cues into account.

In this same vein, it was argued in chapter 2 that the simultaneous nature of trait prediction and the presence of jointly developed features in the shared layer of the multi-task models brings this flavor of automated speaker profiling more into line with how humans likely perform the task of profiling unseen speakers than has been the case for previous, single-task approaches to ASP. However, it may be worth considering the fact that while the MTL models develop representations of all social traits jointly, this is not necessarily the manner in which humans acquire the socio-indexical markers which are used to distinguish between trait categories. Some work examining the acquisition of sociolinguistic variation has hypothesized that, rather than developing representations of all social traits/categories simultaneously, cognitive representations of social traits (and thus representations of the sociolinguistic markers associated with these traits) which are more transparent and/or frequently encountered may be developed earlier in life than representations of those traits which are more complex or less frequently encountered (Foulkes, 2010, pgs. 19-20). Under this hypothesis, socio-indexical markers associated with "easier" category differentiation tasks like speaker sex would therefore be acquired prior to markers associated with potentially more problematic social traits such as ethnicity. Put in terms of computational modeling, this sounds quite similar to work being done on transfer learning, in which models are first trained on one task and then either cross-trained or used as inputs for subsequent models to perform related yet (potentially) harder tasks. Such a transfer learning approach applied to the training procedure of a multi-task model would therefore bring automated speaker profiling even closer to the manner in which humans are hypothesized to perform speaker

[4]E.g. the marked phrase-final rise-and-sustain and rise-and-fall intonation pattern common to Chicano English speakers as described by Fought (2003).

profiling tasks, and may be worth investigating in future work on ASP.[5] That said, the regularization effects of such approaches are precisely what the addition of task-specific output layer weights adjusted during hyper-parameter optimization are designed to achieve in the MTL models presented here, and I'm somewhat skeptical that such a design would outperform the current methodology significantly.

Finally, it could be worthwhile to dig a bit more into the difference or lack-thereof between the STL and MTL models for the region prediction task. As tables 9.1 and 9.2 show, the best performing MTL models outperformed the STL models by at least a small margin in all other trait prediction tasks, but the STL models and the best performing MTL models performed essentially identically on the region prediction task. One possible explanation for this may lie in the nature of the MTL models themselves. As mentioned in chapter 2, MTL is expected to outperform STL only in cases where the primary and secondary model tasks are relatively highly related. It may be the case that while traits such as age, sex, ethnicity, etc. operate on a similar social identity "plane," region operates on a different social identity "plane," and that consequently the linguistic performance of regional identity is less affected by the other four traits considered than was assumed here. In other words, perhaps the other four social traits examined in this dissertation are simply not traits which tend to have a large impact on the linguistic features that the regional prediction task relies on, and consequently the regional prediction task head of the MTL model was not overly involved in the development of the kinds of jointly constructed features within the shared hidden layers that multi-task learning is hypothesized to benefit from. Perhaps we would have seen more of a difference between MTL and STL models aimed at predicting region if we were to include as secondary tasks other social traits that operate on a more similar

[5]Though note that designing ASP systems such that they approximate human behavior simply for the sake of approximating human behavior is irrelevant to the goals of ASP. Human-like implementations should be explored and adopted only insofar as they improve or are hypothesized to improve model performance.

identity "plane" to that of region. Possibilities for this might include secondary tasks for predicting urban vs. rural speakers, a speaker's political orientation, and so on. Further exploration of secondary prediction tasks more closely related to regional origin may prove fruitful for those attempting to model region in a multi-task automated speaker profiling context.

A second possible explanation for the similar performance of the STL and MTL models for region may be that one or more of the other four included social trait prediction tasks acted as a confounding influence on the regional prediction task in the MTL models, and that if one were to remove the offending secondary task(s), one would see an increase in performance for the MTL vs. the STL models. If this were the case, one might also expect that region in turn could have acted as a confounding task for one or more of the other social traits examined during the course of this dissertation, and that removing it would increase performance of the MTL model(s) aimed at predicting those tasks. Ablation experiments in which the MTL models are trained with all possible combinations of the five traits examined would help to clarify whether or not this was in fact the case, and may have the added benefit of elucidating which secondary tasks are most beneficial to which primary tasks– a direction of investigation not explored here, but one which could have practical implications for the deployment of automated speaker profiling systems taking advantage of a multi-task learning framework.

Appendix A: Sinlge-Task Learning Baseline Model Specs

All baseline models were trained using using the functional API of Keras 2.2.4 with a TensorFlow (GPU) 1.13.1 back-end. The TPE algorithm from Hyperopt 0.1.2 was used to perform the hyper parameter search. All training took place using Python 3.6.6. Final tuned hyper parameters are presented in tables A.1 through A.5 for each baseline model below. All unreported parameters were left set to default values.

Table A.1: Tuned hyper parameters for final sex STL models

Parameter	Value
epochs	80
batch size	128
dropout rate	0.44
kernel initialization	glorot_normal
optimizer	rmsprop
hidden Layer 1 size	270
hidden layer 2 size	471

Table A.2: Tuned hyper parameters for final ethnicity STL models

Parameter	Value
epochs	38
batch size	32
dropout rate	0.82
kernel initialization	he_normal
optimizer	adam
hidden Layer 1 size	453
hidden layer 2 size	289

Table A.3: Tuned hyper parameters for final age STL models

Parameter	Value
epochs	49
batch size	128
dropout rate	0.61
kernel initialization	he_uniform
optimizer	adam
hidden Layer 1 size	259
hidden layer 2 size	274

Table A.4: Tuned hyper parameters for final region STL models

Parameter	Value
epochs	20
batch size	128
dropout rate	0.71
kernel initialization	glorot_normal
optimizer	adam
hidden Layer 1 size	323
hidden layer 2 size	577

Table A.5: Tuned hyper parameters for final education STL models

Parameter	Value
epochs	70
batch size	128
dropout rate	0.85
kernel initialization	he_uniform
optimizer	adam
hidden Layer 1 size	612
hidden layer 2 size	173

Appendix B: Multi-Task Learning Model Specs

All MTL models were trained using using the functional API of Keras 2.2.4 with a TensorFlow (GPU) 1.13.1 back-end. The TPE algorithm from Hyperopt 0.1.2 was used to perform the hyper parameter search. All training took place using Python 3.6.6. Final tuned hyper parameters are presented in tables B.1 through B.5 for each baseline model below. All unreported parameters were left set to default values.

Table B.1: Tuned hyper parameters for final sex MTL models

parameter	value
epochs	52
batch_size	128
drop_rate	0.14
init_mode	glorot_normal
optimizer	adam
shared_neurons	21
age_specific_neurons	121
edu_specific_neurons	463
eth_specific_neurons	361
reg_specific_neurons	216
sex_specific_neurons	188
age_weight	0.89
edu_weight	0.61
eth_weight	0.55
reg_weight	0.61
sex_weight	1

Table B.2: Tuned hyper parameters for final ethnicity MTL models

parameter	value
epochs	26
batch_size	128
drop_rate	0.78
init_mode	glorot_normal
optimizer	adam
shared_neurons	632
age_specific_neurons	426
edu_specific_neurons	298
eth_specific_neurons	103
reg_specific_neurons	442
sex_specific_neurons	113
age_weight	0.24
edu_weight	0.47
eth_weight	1
reg_weight	0.36
sex_weight	0.98

Table B.3: Tuned hyper parameters for final age MTL models

parameter	value
epochs	24
batch_size	128
drop_rate	0.79
init_mode	he_uniform
optimizer	adam
shared_neurons	454
age_specific_neurons	183
edu_specific_neurons	456
eth_specific_neurons	199
reg_specific_neurons	107
sex_specific_neurons	107
age_weight	1
edu_weight	0.44
eth_weight	0.28
reg_weight	0.12
sex_weight	0.8

Table B.4: Tuned hyper parameters for final region MTL models

parameter	value
epochs	24
batch_size	128
drop_rate	0.76
init_mode	he_normal
optimizer	adam
shared_neurons	328
age_specific_neurons	415
edu_specific_neurons	95
eth_specific_neurons	110
reg_specific_neurons	184
sex_specific_neurons	154
age_weight	0.67
edu_weight	0.62
eth_weight	0.3
reg_weight	1
sex_weight	0.14

Table B.5: Tuned hyper parameters for final education MTL models

parameter	value
epochs	24
batch_size	128
drop_rate	0.81
init_mode	glorot_normal
optimizer	adam
shared_neurons	622
age_specific_neurons	29
edu_specific_neurons	38
eth_specific_neurons	498
reg_specific_neurons	451
sex_specific_neurons	456
age_weight	0.17
edu_weight	1
eth_weight	0.1
reg_weight	0.77
sex_weight	0.28

Appendix C: Linear Mixed Effects Modeling

Table C.1 presents effect size (η^2, noted in the table as "eta2") and significance information (p-value) for each of the non-ngram individual features examined in chapter 4 with respect to each of the five social traits examined. All models included a fixed effect for the social trait in question, and a random intercept for subject ID.[1] Fixed effects model training and calculation of effect size and p-value were accomplished via afex 0.24 in R.

Only the effect sizes are presented numerically in table C.1. P-values are indicated via asterisks. Those effects reaching significance at $p < 0.001$ are noted with "***", those reaching significance at $p < 0.01$ are noted with "**", and those reaching significance at $p < 0.05$ are noted with "*". The table is not sorted with respect to effect size or significance. Rather, the order of features follows the order in which these features were presented in chapter 4.

Table C.1: Feature effect sizes and significance

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
mean_pitch	0.574***	0.006	0.003	0.018*	0.007
max_pitch	0.016***	0.013	0.026**	0.001	0.005
min_pitch	0.023***	0.001	0.002	0.004	0.012*
jit_loc	0.346***	0.002	0.017*	0.01	0.003
jit_rap	0.296***	0.008	0.014	0.016*	0.001
jit_ppq5	0.405***	0.004	0.008	0.014*	0
shim_loc	0.321***	0.001	0.041***	0.023**	0.002

[1]For example, the formula for the model examining impact of HNR on sex in lmer notation was: HNR ~ Sex + (1|subj_id)

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
shim_apq3	0.212***	0.001	0.034***	0.011	0.001
hnr	0.208***	0.007	0.004	0.012	0.002
vspace_area	0.161***	0.018*	0.015*	0.003	0.002
vspace_dispersion	0.195***	0.022**	0.016*	0.004	0.001
vowel_dynamicity	0	0.006	0.01	0.018*	0.009
f1_25_mean	0.384***	0.005	0.018*	0.016*	0.019**
f1_50_mean	0.386***	0.003	0.013	0.018*	0.017**
f1_75_mean	0.326***	0.001	0.014	0.023**	0.016**
f2_25_mean	0.002	0.005	0.02*	0.004	0.01*
f2_50_mean	0.004	0.007	0.018*	0.007	0.01
f2_75_mean	0.009*	0.005	0.017*	0.008	0.01
f1_25_lob_mean	0	0	0	0	0
f1_50_lob_mean	0	0	0	0	0
f1_75_lob_mean	0	0	0	0	0
f2_25_lob_mean	0	0	0	0	0
f2_50_lob_mean	0	0	0	0	0
f2_75_lob_mean	0	0	0	0	0
f1_25_range	0.009*	0.011	0.013	0.005	0.001
f1_50_range	0.002	0.017*	0.014	0.002	0.001
f1_75_range	0.004	0.008	0.009	0.002	0.002
f2_25_range	0.087***	0.005	0.016*	0.003	0.004
f2_50_range	0.117***	0.006	0.011	0.005	0
f2_75_range	0.114***	0.004	0.02*	0.008	0.001
f1_25_lob_range	0.146***	0.019*	0.014	0.012	0.006
f1_50_lob_range	0.112***	0.009	0.031***	0.008	0.005
f1_75_lob_range	0.111***	0.003	0.027**	0.008	0.005
f2_25_lob_range	0.049***	0.025**	0.007	0.006	0.028***
f2_50_lob_range	0.1***	0.041***	0.006	0.011	0.011*
f2_75_lob_range	0.073***	0.011	0.008	0.016*	0.008
AE1_f1	0.083***	0.07***	0.024**	0.044***	0.006
AO1_f1	0.009*	0.01	0.004	0.037***	0.002
UH1_f1	0.01*	0.008	0.009	0.005	0.003
OY1_f1	0.01	0.013	0.076	0.088	0.014
UW1_f1	0.014**	0.006	0	0.005	0.011*
IH1_f1	0.007*	0.026**	0.01	0.017*	0.003
EH1_f1	0.024***	0.154***	0.031***	0.041***	0.042***
OW1_f1	0.009*	0.025**	0.012	0.019**	0.014*
AW1_f1	0.038***	0.026**	0.011	0.002	0.004

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
EY1_f1	0.011**	0.013	0.006	0.012	0.007
AH1_f1	0	0.009	0.02*	0.006	0.005
AY1_f1	0.001	0.128***	0.173***	0.044***	0.029***
ER1_f1	0	0.004	0.018	0.055***	0
AA1_f1	0.005	0.028***	0.003	0.005	0.004
IY1_f1	0.004	0.012	0.005	0.009	0
AE1_f2	0.16***	0.056***	0.007	0.029***	0.003
AO1_f2	0.114***	0.009	0.015	0.021**	0.001
UH1_f2	0.039***	0.004	0.002	0.013	0.001
OY1_f2	0.079*	0.048	0.031	0.036	0.041
UW1_f2	0.024***	0.011	0.005	0.016*	0
IH1_f2	0.147***	0.017*	0.003	0.001	0.008
EH1_f2	0	0.056***	0.026**	0.019**	0.001
OW1_f2	0.181***	0.033***	0.029***	0.027**	0.003
AW1_f2	0.192***	0.021**	0.023**	0.006	0.001
EY1_f2	0.33***	0.006	0.004	0.014*	0.002
AH1_f2	0.336***	0.007	0.008	0.038***	0.001
AY1_f2	0.074***	0.019*	0.012	0.005	0.017**
ER1_f2	0.152***	0.032**	0.048***	0.017*	0.004
AA1_f2	0.253***	0.003	0.001	0.018*	0.001
IY1_f2	0.366***	0.005	0.002	0.016*	0.003
AH1_f1_25_lob	0.001	0.004	0.021**	0.009	0.001
AH1_f1_50_lob	0	0.009	0.02*	0.006	0.005
AH1_f1_75_lob	0	0.005	0.014	0.007	0.003
AH1_f2_25_lob	0.304***	0.011	0.004	0.039***	0.004
AH1_f2_50_lob	0.336***	0.007	0.008	0.038***	0.001
AH1_f2_75_lob	0.377***	0.009	0.007	0.03***	0.001
AA1_f1_25_lob	0.006*	0.034***	0.002	0.011	0.002
AA1_f1_50_lob	0.005	0.028***	0.003	0.005	0.004
AA1_f1_75_lob	0.026***	0.022**	0.003	0.003	0.003
AA1_f2_25_lob	0.203***	0.004	0.001	0.019*	0.002
AA1_f2_50_lob	0.253***	0.003	0.001	0.018*	0.001
AA1_f2_75_lob	0.316***	0.003	0.005	0.014*	0.002
AY1_f1_25_lob	0.012**	0.08***	0.178***	0.029***	0.016**
AY1_f1_50_lob	0.001	0.128***	0.173***	0.044***	0.029***
AY1_f1_75_lob	0.004	0.161***	0.079***	0.035***	0.026***
AY1_f2_25_lob	0.242***	0.014*	0.007	0.014*	0.005
AY1_f2_50_lob	0.074***	0.019*	0.012	0.005	0.017**

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
AY1_f2_75_lob	0.054***	0.017*	0.009	0.02**	0.006
UW1_f1_25_lob	0.023***	0.007	0.002	0.013*	0.008
UW1_f1_50_lob	0.014**	0.006	0	0.005	0.011*
UW1_f1_75_lob	0.045***	0.003	0.003	0.005	0.005
UW1_f2_25_lob	0.062***	0.01	0.005	0.013	0
UW1_f2_50_lob	0.024***	0.011	0.005	0.016*	0
UW1_f2_75_lob	0.004	0.008	0.001	0.009	0
EY2_f1_25_lob	0.002	0.049	0.054	0.014	0.06*
EY2_f1_50_lob	0	0.006	0.028	0.033	0.061*
EY2_f1_75_lob	0.028	0.013	0.033	0.043	0.036
EY2_f2_25_lob	0.051*	0.011	0.048	0.049	0.009
EY2_f2_50_lob	0.072**	0.006	0.069	0.116**	0
EY2_f2_75_lob	0.127***	0.008	0.045	0.043	0.008
OW1_f1_25_lob	0.01**	0.035***	0.066***	0.033***	0.023***
OW1_f1_50_lob	0.009*	0.025**	0.012	0.019**	0.014*
OW1_f1_75_lob	0.047***	0.002	0.003	0.012	0.005
OW1_f2_25_lob	0.256***	0.057***	0.01	0.02**	0.004
OW1_f2_50_lob	0.181***	0.033***	0.029***	0.027**	0.003
OW1_f2_75_lob	0.135***	0.027**	0.033***	0.019**	0.002
AO1_f1_25_lob	0.019***	0.002	0.001	0.043***	0.001
AO1_f1_50_lob	0.009*	0.01	0.004	0.037***	0.002
AO1_f1_75_lob	0.004	0.01	0.004	0.025**	0.004
AO1_f2_25_lob	0.071***	0.009	0.014	0.017*	0.002
AO1_f2_50_lob	0.114***	0.009	0.015	0.021**	0.001
AO1_f2_75_lob	0.206***	0.011	0.012	0.019*	0
EH1_f1_25_lob	0.009*	0.138***	0.043***	0.034***	0.04***
EH1_f1_50_lob	0.024***	0.154***	0.031***	0.041***	0.042***
EH1_f1_75_lob	0.055***	0.147***	0.012	0.032***	0.029***
EH1_f2_25_lob	0.018***	0.036***	0.02*	0.007	0.001
EH1_f2_50_lob	0	0.056***	0.026**	0.019**	0.001
EH1_f2_75_lob	0.012**	0.045***	0.01	0.023**	0.006
EY1_f1_25_lob	0	0.059***	0.078***	0.027**	0.022**
EY1_f1_50_lob	0.011**	0.013	0.006	0.012	0.007
EY1_f1_75_lob	0.022***	0.006	0.018*	0.004	0
EY1_f2_25_lob	0.272***	0.008	0.006	0.008	0.001
EY1_f2_50_lob	0.33***	0.006	0.004	0.014*	0.002
EY1_f2_75_lob	0.377***	0.008	0.006	0.017*	0.002
AW1_f1_25_lob	0.034***	0.008	0.011	0.006	0.003

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
AW1_f1_50_lob	0.038***	0.026**	0.011	0.002	0.004
AW1_f1_75_lob	0.019***	0.036***	0.012	0.005	0.004
AW1_f2_25_lob	0.111***	0.037***	0.024**	0.014*	0.004
AW1_f2_50_lob	0.192***	0.021**	0.023**	0.006	0.001
AW1_f2_75_lob	0.178***	0.004	0.013	0.004	0.001
AA2_f1_25_lob	0.001	0.072	0.054	0.084*	0.002
AA2_f1_50_lob	0.008	0.052	0.052	0.116*	0.016
AA2_f1_75_lob	0.004	0.041	0.033	0.06	0.013
AA2_f2_25_lob	0.165***	0.013	0.066	0.025	0.006
AA2_f2_50_lob	0.155***	0.011	0.069	0.007	0.041
AA2_f2_75_lob	0.173***	0.012	0.101*	0.011	0.032
IH1_f1_25_lob	0.068***	0.013	0.032***	0.011	0.005
IH1_f1_50_lob	0.007*	0.026**	0.01	0.017*	0.003
IH1_f1_75_lob	0.002	0.032***	0.011	0.01	0.006
IH1_f2_25_lob	0.12***	0.031***	0.003	0.003	0.006
IH1_f2_50_lob	0.147***	0.017*	0.003	0.001	0.008
IH1_f2_75_lob	0.139***	0.025**	0.015	0.002	0.004
ER1_f1_25_lob	0	0.032**	0.015	0.062***	0.004
ER1_f1_50_lob	0	0.004	0.018	0.055***	0
ER1_f1_75_lob	0.016**	0.013	0.01	0.017*	0.003
ER1_f2_25_lob	0.116***	0.023*	0.038***	0.008	0.003
ER1_f2_50_lob	0.152***	0.032**	0.048***	0.017*	0.004
ER1_f2_75_lob	0.132***	0.016	0.018	0.008	0.002
UH1_f1_25_lob	0.034***	0.003	0.018*	0	0.003
UH1_f1_50_lob	0.01*	0.008	0.009	0.005	0.003
UH1_f1_75_lob	0	0.008	0.011	0.009	0.001
UH1_f2_25_lob	0.026***	0.004	0.006	0.01	0.004
UH1_f2_50_lob	0.039***	0.004	0.002	0.013	0.001
UH1_f2_75_lob	0.068***	0.01	0.011	0.019*	0.001
AE1_f1_25_lob	0.083***	0.053***	0.022**	0.038***	0.004
AE1_f1_50_lob	0.083***	0.07***	0.024**	0.044***	0.006
AE1_f1_75_lob	0.093***	0.061***	0.011	0.034***	0.006
AE1_f2_25_lob	0.274***	0.041***	0.003	0.014*	0.002
AE1_f2_50_lob	0.16***	0.056***	0.007	0.029***	0.003
AE1_f2_75_lob	0.082***	0.06***	0.008	0.031***	0.004
IY1_f1_25_lob	0.026***	0.009	0.003	0.01	0.002
IY1_f1_50_lob	0.004	0.012	0.005	0.009	0
IY1_f1_75_lob	0	0.012	0.014	0.013*	0.001

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
IY1_f2_25_lob	0.347***	0.005	0.004	0.016*	0.002
IY1_f2_50_lob	0.366***	0.005	0.002	0.016*	0.003
IY1_f2_75_lob	0.353***	0.003	0.005	0.009	0.005
ER2_f1_25_lob	0.09	0.107	0.217	0.218	0.188
ER2_f1_50_lob	0.04	0.063	0.422	0.112	0.058
ER2_f1_75_lob	0.032	0.021	0.626*	0.172	0.028
ER2_f2_25_lob	0.134	0.032	0.427	0.211	0.163
ER2_f2_50_lob	0.318*	0.024	0.305	0.217	0.256
ER2_f2_75_lob	0.317*	0.087	0.142	0.165	0.434
UW2_f1_25_lob	0.009	0.039	0.02	0.033	0.008
UW2_f1_50_lob	0.028	0.042	0.049	0.066	0.009
UW2_f1_75_lob	0.004	0.051	0.103*	0.094*	0.013
UW2_f2_25_lob	0.003	0.031	0.014	0.016	0.069*
UW2_f2_50_lob	0.002	0.008	0.024	0.042	0.022
UW2_f2_75_lob	0.01	0.005	0.008	0.042	0.004
EH2_f1_25_lob	0.009	0.029	0.093**	0.053	0.003
EH2_f1_50_lob	0.001	0.016	0.04	0.012	0.007
EH2_f1_75_lob	0.003	0.026	0.015	0.009	0.004
EH2_f2_25_lob	0	0.026	0.009	0.005	0.004
EH2_f2_50_lob	0	0.058	0.014	0.006	0.022
EH2_f2_75_lob	0	0.067*	0.005	0.009	0.025
IY2_f1_25_lob	0.014	0.004	0.025	0.042	0.068*
IY2_f1_50_lob	0	0.058	0.151**	0.049	0.055
IY2_f1_75_lob	0	0.141**	0.084	0.038	0.081*
IY2_f2_25_lob	0.211***	0.004	0.046	0.017	0.01
IY2_f2_50_lob	0.187***	0.04	0.021	0.006	0.002
IY2_f2_75_lob	0.129***	0.054	0.034	0.072	0.03
OY1_f1_25_lob	0.053	0.013	0.074	0.097	0.038
OY1_f1_50_lob	0.01	0.013	0.076	0.088	0.014
OY1_f1_75_lob	0	0.044	0.039	0.091	0.003
OY1_f2_25_lob	0.039	0.062	0.129	0.038	0.017
OY1_f2_50_lob	0.079*	0.048	0.031	0.036	0.041
OY1_f2_75_lob	0.03	0.072	0.038	0.019	0.012
OW2_f1_25_lob	0.003	0.009	0.017	0.02	0.002
OW2_f1_50_lob	0.016	0.007	0.011	0.025	0.01
OW2_f1_75_lob	0.006	0.002	0.012	0.026	0.013
OW2_f2_25_lob	0.062***	0.02	0.074**	0.01	0.009
OW2_f2_50_lob	0.067***	0.009	0.053*	0.013	0.018

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
OW2_f2_75_lob	0.05**	0.009	0.035	0.007	0.012
IH2_f1_25_lob	0.039***	0.011	0.038*	0.004	0.025*
IH2_f1_50_lob	0.002	0.016	0.019	0.008	0.001
IH2_f1_75_lob	0	0.015	0.026	0.007	0.004
IH2_f2_25_lob	0.028**	0.01	0.003	0.011	0.001
IH2_f2_50_lob	0.01	0.005	0.022	0.029	0.001
IH2_f2_75_lob	0.029**	0.005	0.024	0.021	0.003
say	0.001	0.023**	0.012	0.019**	0.006
go	0	0.002	0.003	0.004	0.014*
be_like	0.001	0.006	0.137***	0.008	0.005
be_all	0.002	0.006	0.009	0.003	0.001
really	0.001	0.021**	0.063***	0.022**	0.022**
like	0	0.018*	0.239***	0.036***	0.012*
just	0.005	0.013	0.011	0.001	0.001
well	0	0.002	0.008	0.007	0.007
okay	0.001	0.059***	0.007	0.005	0.007
yeah	0.014**	0.015*	0.019*	0.013*	0.009
right	0.005	0.006	0.011	0.009	0.002
so	0.002	0.055***	0.021**	0.019*	0.001
kinda	0.001	0.001	0.01	0.007	0
sorta	0.002	0.004	0.01	0.003	0.007
kind_of	0	0.012	0.017*	0.004	0.021**
sort_of	0.002	0.017*	0.006	0.004	0.012*
you_know	0.001	0.025**	0.04***	0.005	0.014*
i_mean	0	0.001	0.011	0.001	0.002
i_guess	0.004	0.005	0.02*	0.008	0.009
i_know	0.026***	0.006	0.015*	0.007	0
will	0.003	0.004	0.008	0.008	0.007
would	0.002	0.018*	0.008	0.003	0.006
shall	0.002	0.002	0.001	0.002	0.011*
should	0.002	0.009	0.004	0.006	0.014*
may	0	0	0.011	0.003	0
might	0.004	0.002	0.01	0.005	0.005
can	0.001	0.002	0.003	0.005	0.004
could	0.002	0.015*	0.01	0.007	0.003
ought	0.002	0.015*	0.009	0.009	0.002
must	0.005	0.006	0.004	0.005	0.013*
going_to	0	0	0	0	0

Table C.1: Feature effect sizes and significance (cont.)

feature	eta2_sex	eta2_eth	eta2_age	eta2_reg	eta2_edu
have_to	0.004	0.009	0.015*	0.008	0.009
need_to	0	0.004	0.005	0.005	0.01
pron_1st_prop	0.001	0.034***	0.043***	0.008	0.007
pron_2nd_prop	0.005	0.027**	0.004	0.003	0.002
pron_3rd_prop	0.002	0.022**	0.046***	0.016*	0.005
i	0.009*	0.005	0.019*	0.007	0.001
taboo_freq	0	0.002	0.013	0.003	0.001
polite_freq	0.005	0.002	0.009	0.005	0.002
intense_very	0.006*	0.019*	0.016*	0.005	0
intense_so	0.007*	0.008	0.003	0.008	0
intense_really	0.004	0.029***	0.039***	0.021**	0.023***
intense_too	0	0.002	0.012	0.002	0
intense_real	0	0.013	0.014	0.008	0.009
intense_right	0.003	0.01	0.006	0.003	0.002
intense_pretty	0.019***	0.016*	0.035***	0.012	0.013*
intense_totally	0.003	0.005	0.02*	0.002	0.002
intense_completely	0	0.005	0.013	0.001	0.002
intense_absolutely	0	0.003	0.018*	0.009	0.001
intense_highly	0	0	0	0	0
intense_seriously	0.001	0.001	0.011	0.014*	0.003
intense_damn	0.001	0.001	0.013	0.003	0.007
intense_fucking	0	0	0	0	0
avg_syl_len	0	0.017*	0.017*	0.019*	0.029***
tok_per_min	0.001	0.018*	0.016*	0.014*	0.003
sent_polarity	0.002	0.018*	0.005	0.004	0.006
sent_subjectivity	0.001	0.011	0.008	0.014*	0.002
quotative_freq	0.001	0.003	0.05***	0.002	0.009
modal_freq	0.003	0.017*	0.013	0	0.013*
discourse_marker_freq	0.006*	0.035***	0.098***	0.025**	0.005
intensifier_freq	0.002	0.035***	0.016*	0.019**	0.008

Appendix D: Lexical Sets

The full set of lexical items belonging to each group of lexical features discussed in chapter 4 are listed below.

D.1 Quotatives

Drawn from Barbieri (2008):

- Go
- Say
- Be like
- Be all

D.2 Modals

Drawn from Barbieri (2008):

- will
- would
- shall
- should
- may
- might
- can
- could
- ought
- must
- going to
- have to
- need to

D.3 Discourse Markers

Drawn from Barbieri (2008):

- really
- like
- just
- well
- okay
- yeah
- right
- so
- kinda
- sorta
- kind of
- sort of
- you know
- i mean
- i guess
- i know

D.4 Politeness Markers

Drawn from Biber et al. (1999):

- please
- thank
- thanks
- sorry
- pardon
- excuse

D.5 Taboo Markers

Drawn from Lancker and Cummings (1999):

- fuck
- fucking
- cunt
- shit
- bastard
- damn
- goddamn
- god
- whore
- nigger
- fascist
- hell
- heck
- crap
- bitch
- christ
- screw
- piss
- ass
- butt
- slut
- ream
- shaft
- balls
- jesus
- cock
- prick
- dick
- penis
- fart
- asshole
- bullshit
- blow
- blowjob
- dildo
- motherfucker
- queer
- spic
- dipshit

D.6 Intensifiers

Drawn from Barbieri (2008):

- very
- so
- really
- too
- real
- right
- pretty
- totally
- completely
- absolutely
- highly
- seriously
- damn
- fucking

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283.

Alsharhan, E. and Ramsay, A. (2019). Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions. *Information Processing & Management*, 56(2):343–353.

Anderson, K. and Leaper, C. (1998). Emotion talk between same- and mixed- gender friends: Form and function. *Journal of Language and Sexuality*, 17(4):419–448.

Ardehaly, E. M. and Culotta, A. (2015). Inferring latent attributes of Twitter users with label regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–195.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, 18(2):135–160.

Baranowski, M. A. (2008). The fronting of the back upgliding vowels in Charleston, South Carolina. *Language Variation and Change*, 20(03):527.

Barbieri, F. (2007). Older men and younger women: A corpus-based study of quotative use in American English. *English World-Wide*, 28(1):23–45.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics*, 12(1):58–88.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

Bauman, C. (2014). "Oh [o:], I'm the token Asian": A potential vowel marker of ethnic identity. *43rd Annual Conference on New Ways of Analyzing Variation (NWAV 43)*.

Bayard, D. and Krishnayya, S. (2001). Gender, expletive use, and context: Male and female expletive use in structured and unstructured conversation among New Zealand university students. *Women and Language*, 24(1):1–15.

Bayley, R. and Santa Ana, O. (2004). Chicano English: Morphology and syntax. In Kortmann, B., Schneider, E. W., Upton, C., Mesthrie, R., and Burridge, K., editors, *A Handbook of Varieties of English vol. 2: Morphology and Syntax*, pages 374–390. Mouton de Gruyter, Berlin.

Bell, A. (1984). Language style as audience design. *Language in Society*, 13(2):145–201.

Benor, S. B. (2010). Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity. *Journal of Sociolinguistics*, 14(2):159–183.

Bergstra, J., Bardenet, R., Bengio, Y., and Kegl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24*, pages 2546–2554.

Bergstra, J., Yamins, D. L. K., and Cox, D. (2013). Making a science of model search:

Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123.

Biadsy, F. (2011). *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*. PhD thesis, Columbia University.

Biadsy, F., Hirschberg, J., and Ellis, D. P. (2011). Dialect and accent recognition using phonetic-segmentation supervectors. In *INTERSPEECH*, pages 745–748.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education Limited, Edinburgh.

Blyth, C., Recktenwald, S., and Wang, J. (1990). "I'm like, 'Say what?!': A new quotative in American oral narrative". *American Speech*, 65:215–227.

Bocklet, T., Noth, E., Stemmer, G., Ruzickova, H., and Rusz, J. (2011). Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis. In *Automatic Speech Recognition and Understanding (ASRU)*, pages 478–483.

Boulis, C. and Ostendorf, M. (2005). A quantitative analysis of lexical differences between genders in telephone conversations. *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, (June):435–442.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, L. (2011). Sibilant variation and gender identity. In *NWAV 40, Georgetown University*, Washington, D.C.

Brown, L. (2015). *Phonetic Cues and the Perception of Gender and Sexual Orientation*. PhD thesis, University of Toronto.

Bucholtz, M. (1999). "Why Be Normal?": Language and identity practices in a community of Nerd Girls. *Language in Society*, 28(2):203–223.

Bucholtz, M. (2004). Styles and stereotypes: The linguistic negotiation of identity among Laotian American youth. *Pragmatics*, 14(2-3):127–147.

Bucholtz, M. (2009). From Stance to Style: Gender, Interaction, and Indexicality in Mexican Immigrant Youth Slang. In *Stance: Sociolinguistic Perspectives*.

Burkhardt, F., Eckert, M., Johannsen, W., and Stegmann, J. (2010). A database of age and gender annotated telephone speech. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1562–1565, Valletta, Malta.

Callahan-Price, E. E. (2013). *Emerging Hispanic English in the Southeast U.S.: Grammatical Variation in a Triethnic Community*. PhD thesis, Duke University.

Campbell, H., Bell, M. M., and Finney, M. (2006). *Country boys: Masculinity and rural life*. Penn State Press.

Campbell-Kibler, K. (2006). *Listener Perceptions of Sociolinguistic Variables: The Case of (ING)*. PhD thesis, Stanford University.

Caruna, R. (1997). Multitask learning. *Machine Learning*, 28:41–75.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder. *arXiv preprint*.

Chambers, J. (2003). *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Blackwell, Oxford.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet. *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768.

Cheshire, J. (2005). Syntactic variation and beyond: gender and social class variation in the use of discourse-new markers. *Journal of Sociolinguistics*, 9:479–507.

Chollet, F. (2015). Keras. *https://keras.io*.

Chun, E. W. (2001). The construction of White, Black, and Korean American identities through African American Vernacular English. *Journal of Linguistic Anthropology*, 11(1):52–64.

Clyne, M. (2000). Lingua franca and ethnolects in Europe and beyond. *Sociolinguistica*, 14:83–89.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing. *Proceedings of the 25th international conference on Machine learning*, 20(1):160–167.

Coulthard, M., Johnson, A., and Wright, D. (2016). *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge, New York.

Coupland, N. (1980). Style-Shifting in a Cardiff Work-Setting. *Language in Society*, 9(1):1–12.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, pages 139–168.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representatations for monosyllabic word recognition in continuosly spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366.

Doddington, G. (2001). Speaker Recognition based on Idiolectal Differences between Speakers. *Eurospeech*, 4:2517– 2520.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1723–1732.

Eckert, P. (1989). The whole woman: Sex and gender difference in variation. *Language Variation and Change*, 1(3):245–267.

Eckert, P. (1997). Age as a sociolinguistic variable. *The hand book of socialinguistics*, (1990):151–167.

Eckert, P. (2008a). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Eckert, P. (2008b). Where do ethnolects stop? *International Journal of Bilingualism*, 12(1-2):25–42.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41:87–100.

Eckert, P. and McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. *Language in Society*, 28(02):185–201.

Eckert, P. and McConnell-Ginet, S. (2003). *Language & Gender*. Cambridge University Press., Cambridge.

Eddington, D. and Taylor, M. (2009). T-Glottalization in American English. *American Speech*, 84(3):298–314.

Eisenstein, J. (2015). Identifying regional dialects in online social media. In Boberg, C., Nerbonne, J., and Watt, D., editors, *Handbook of Dialectology*. Wiley.

Evgeniou, T. and Pontil, M. (2004). Regularized multi-task learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117.

Ferguson, S. H. and Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research*, 50(5):1241–1255.

Fink, C., Kopecky, J., and Morawski, M. (2012). Inferring gender from the content of tweets: A region specific example. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 459–462.

Fisher, A., Rudin, C., and Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint*, arXiv:1801.

Fisher, S., Prichard, H., and Sneller, B. (2015). The apple doesn't fall far from the tree: Incremental change in Philadelphia families. *Penn Working Papers in Linguistics*, 21(2):49–58.

Fought, C. (2003). *Chicano English in Context*. Palgrave, Basingstoke.

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, 1(1):5–39.

Fridland, V. (2001). The social dimension of the Southern vowel shift: Gender, age and class. *Journal of Sociolinguistics*, 5(2):233–253.

Gillick, D. (2010). Can conversational word usage be used to predict speaker demographics ? In *Proceedings of Interspeech 2010*, pages 1381–1384, Makuhari, Japan.

Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Graddol, D. and Swann, J. (1983). Speaking fundamental frequency: Some physical and social correlates. *Language and Speech1*, 26(4):351–366.

Green, L. (2002). *African American English: A Linguistic Introduction*, volume 35. Cambridge University Press, New York.

Hagan, M. and Demuth, H. (1995). *Neural Network Design*.

Hall-Lew, L. (2005). One shift, two groups: When fronting alone is not enough. *Penn Working Papers in Linguistics*, 10(2):105–116.

Hall-Lew, L. (2009). Ethnic practice is local practice: Phonetic change in San Francisco, California. *Poster presented at VoxCalifornia: Cultural meanings of linguistic Diversity*.

Hall-Lew, L. (2011). The completion of a sound change in California English. *17th International Congress of Phonetic Sciences (ICPhS XVII)*, pages 807–810.

Hazenberg, E. (2012). *Language and Identity Practice: A sociolinguistic study of gender in Ottowa, Ontario*. PhD thesis, Memorial University of Newfoundland.

Hill, D. R. (2007). Speaker classification concepts: Past, present and future. In Müller, C., editor, *Speaker Classification: Fundamentals, Features, and Methods*, chapter 2, pages 21–46. Springer-Verlag, Berlin.

Hoffman, M. F. and Walker, J. A. (2010). Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change*, 22:37–67.

Hooks, B. (1981). *Ain't I am Woman? Black Women and Feminism*. South End Press, Boston.

Hu, Y., Wu, D., and Nucci, A. (2012). Pitch-based gender identification with two-stage classification. *Security and Communication Networks*, 5(2):211–225.

Humes, K. R., Jones, N. a., and Ramirez, R. R. (2011). Overview of race and hispanic origin: 2010. *2010 Census Briefs*.

Ito, R. and Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, 32(2):257–279.

Jacewicz, E., Fox, R. A., and Lyle, S. (2009). Variation in stop consonant voicing in two regional varieties of American English. *The Journal of the International Phonetic Association*, 39(3):313–334.

Jacewicz, E., Fox, R. A., and Salmons, J. (2011). Vowel change across three age groups of speakers in three regional varieties of American English. *Journal of Phonetics*, 39(4):683–693.

Jaspers, J. (2008). Problematizing ethnolects: Naming linguistic practices in an Antwerp secondary school. *International Journal of Bilingualism*, 12(1/2):85–103.

Jessen, M. (2007). Speaker Classification in Forensic Phonetics and Acoustics. In Müller, C., editor, *Speaker Classification: Fundamentals, Featurs, and Methods*, chapter 10, pages 180–204. Springer-Verlag, Berlin.

Johannsen, A., Hovy, D., and Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL)*, pages 103–112.

Johnstone, B., Andrus, J., and Danielson, A. (2006). Mobility, indexicality, and the enregisterment of "Pittsburghese". *Journal of English Linguistics*, 34(2):77–104.

Johnstone, B. and Kiesling, S. F. (2008). Indexicality and Experience: Variation and Identity in Pittsburgh. *Journal of Sociolinguistics*, 12(1):5–33.

Kamaruddin, N., Rahman, A. W. A., and Shah, A. N. R. (2016). Measuring customer satisfaction through speech using valence-arousal approach. In *6th International*

Conference on Information and Communication Technology for The Muslim World (ICT4M), pages 298–303.

Keating, P., Garellek, M., and Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *ICPhS 2015*, (1):2–7.

Kendall, T. (2013). *Speech Rate, Pause and Sociolinguistic Variation: Studies in corpus sociophonetics*. Palgrave Macmillan, New York.

Kennedy, R. and Grama, J. (2012). Chain Shifting and Centralization in California Vowels: An Acoustic Analysis. *American Speech*, 87(1):39–56.

Kiesling, S. F. (2004). Dude. *American Speech*, 79(3):281–305.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40.

Kirkham, S. (2015). Intersectionality and the social meanings of variation: Class, ethnicity, and social practice. *Language in Society*, 44(5):629–652.

Kirtley, M. J., Grama, J., Drager, K., and Simpson, S. (2016). An acoustic analysis of the vowels of Hawai‘i English. *Journal of the International Phonetic Association*, 46(01):79–97.

Kohn, M. E. (2008). *Latino English in North Carolina: A Comparison of Emerging Communities*. PhD thesis, North Carolina State University.

Kuhn, M. C. f. J. W., Weston, S., Williams, A., Keefer, C., and Engelhardt, A. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5).

Künzel, H. J. (2001). Beware of the ‘ telephone effect ’: the influence of telephone transmission on the measurement of formant frequencies. 8(1):1350–1771.

Labov, W. (1966). *The social stratification of English in New York City*. Center for Applied Linguistics, Washington, D.C.

Labov, W. (1972a). *Language in the inner city: Studies in the Black English Vernacular*. Blackwell, Oxford.

Labov, W. (1972b). *Sociolinguistic Patterns*. Blackwell, Oxford.

Labov, W. (1972c). Sociolinguistic Patterns. *Language*, 2(4):344.

Labov, W. (1978). Where does the linguistic variable stop? A response to Beatriz Lavandera. *Working papers in sociolinguistics*, 44:1–17.

Labov, W. (1994). *Principles of Linguistic Change. Vol.1, Internal factors*. Blackwell, Oxford.

Labov, W. (2001). *Principles of Linguistic Change. Vol.2, Social factors*. Blackwell, Oxford.

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

Lakoff, R. (1975). *Language and Woman's Place*. Harper and Row, New York.

Lancker, D. V. and Cummings, J. L. (1999). Expletives: Neurolinguistic and neurobehavioral inquiries into swearing. *Brain Research Reviews*, 31(1):83–104.

Lavandera, B. R. (1978). Where does the sociolinguistic variable stop? *Language in Society*, 7(2):171–182.

Lee, B. J., Keun Ho, K., Ku, B., Jang, J.-S., and Kim, J. Y. (2013). Prediction of body mass index status from voice signals based on machine learning for automated medical applications. *Artificial Intelligence in Medicine*, 58(1):51–61.

Levitan, S. I., Mishra, T., and Bangalore, S. (2016). Automatic identification of gender from speech. In *Proceedings of Speech Prosody 2016*, pages 84–88.

Levon, E. (2015). Integrating intersectionality in language, gender, and sexuality research. *Language and Linguistics Compass*, 9(7):295–308.

Li, M., Han, K. J., and Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151–167.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Liu, X., Gao, J., He, X., Deng, L., Duh, K., and Wang, Y.-y. (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 912–921, Denver, CO.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different listeners. *Journal of the Acoustic Society of America*, 49(2B):606–608.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner [Computer program].

Mendoza-Denton, N. (1997). *Chicana/Mexicana Identity and Linguistic Variation: An Ethnographic and Sociolinguistic Study of Gang Affiliation in an Urban High School*. PhD thesis, Stanford University.

Mendoza-Denton, N. (2011). Creaky Voice, Circulation, and Gendered Hardcore in a Chicana/o Gang Persona. *Journal of Linguistic Anthropology*, 21(2):260–278.

Milroy, L. (1980). *Language and social networks*. Blackwell, Oxford.

Mondorf, B. (2002). Gender differences in English syntax. *Journal of English Linguistics*, 30(2):158–180.

Moosavi, N. S. and Strube, M. (2017). Lexical Features in Coreference Resolution: To be Used With Caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

Munson, B. (2007). Lexical characteristic mediate the influence of sex and sex typicality on vowel-space size. *Proceedings of International Congress of Phonetic Sciences*, pages 885–888.

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech Language and Hearing Research*, 51(3):574–585.

Newman, M. and Wu, A. (2011). "Do you sound Asian when you speak English?" Racial identification and voice in Chinese and Korean Americans' English. *American Speech*, 86(2):152–178.

Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of the seventh international AAAI conference on weblogs and social media*, pages 439–448, Cambridge, MA.

Parameswaran, S. and Weinberger, K. Q. (2010). Large margin multi-task metric learning. In *Advances in Neural Information Processing Systems (NIPS) 23*, pages 867–1875.

Podesva, R. J. and Kajino, S. (2014). Sociophonetics, Gender, and Sexuality. In Ehrlich, S., Meyerhoff, M., and Holmes, J., editors, *The Handbook of Language, Gender, and Sexuality,*, pages 103–122. Wiley-Blackwell, Malden, MA.

Podesva, R. J. and Van Hofwegen, J. (2014). How Conservatism and Normative Gender Constrain Variation in Inland California : The Case of /s/. *University of Pennsylvania Working Papers in Linguistics*, 20(2):129–137.

Poorjam, A. H., Bahari, M. H., and Hamme, H. V. (2014). Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In *4th International Conference on Computer and Knowledge Engineering (ICCKE)*.

Precht, K. (2008). Sex similarities and differences in stance in informal American conversation. *Journal of Sociolinguistics*, 12(1):89–111.

Prichard, H. and Tamminga, M. (2012). The Impact of Higher Education on Philadelphia Vowels. *U. Penn Working Papers in Linguistics*, 18(2):87–95.

Purnell, T., Idsardi, W., and Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology*, 18(1):10–30.

Purnell, T., Tepeli, D., and Salmons, J. (2005). German substrate effects in Wisconsin English: evidence for final fortition. *American Speech*, 80(2):135–164.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., and Xiang, B. (2003). The SuperSID Project : Exploiting High-level Information for High-accuracy speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pages 784–787, Hong Kong.

Richardson, G. (1984). Can Y'all function as a singular pronoun in Southern dialect? *American Speech*, 59(1):51–59.

Rickford, J. (1999). *African American Vernacular English: Features, evolution, educational implications*. Blackwell, Malden, MA.

Rickford, J. and McNair-Knox, F. (1994). Addressee-and topic-influenced style shift: A quantitative sociolinguistic study. In Biber and Finegan, editors, *Sociolinguistic Perspectives on Register*, pages 235–276. Oxford University Press, Oxford.

Rickford, J. and Price, M. (2013). Girlz II women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics*, 17(2):143–179.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite.

Sankoff, G. and Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 83(3):560–588.

Schilling, N. and Marsters, A. (2015). Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes. *Annual Review of Applied Linguistics*, 35:195–214.

Schleef, E. (2005). Gender, power, discipline, and context: On the sociolinguistic variation of okay, right, like, and you know in English academic discourse. In *Proceedings of the 12th Annual symposium about Language and Society*, pages 177–186, Austin.

Schuller, B., Steidl, S., Batliner, A., Noth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., and Weiss, B. (2012). The INTERSPEECH 2012 Speaker Trait Challenge. *13th Annual Conference of the International Speech Communication Association*, pages 254–257.

Shafran, I., Riley, M., and Mohri, M. (2003). Voice signatures. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 31–36.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, 23(3-4):193–229.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Linguistics and Language Compass*, 3(2):621–640.

Simpson, A. P. and Ericsdotter, C. (2007). Sex-specific differences in F0 and vowel space. In *Proceedings of the16th International Congress of Phonetic Sciences*, pages 933–936.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In Coates and Hualde, I., editors, *Laboratory Phonology 9*, pages 65– 86. Mouton de Gruyter, New York.

Stubbe, M. and Holmes, J. (1995). You know, eh and other 'exasperating expressions': An analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language and Communication*, 15(1):63–88.

Sulayes, R. (2009). *An experiment in automated linguistic profiling of transcribed speech*. Unpublished Manuscript, Washington, D.C.

Tagliamonte, S. and D'Arcy, A. (2004). He's like, she's like: The quotative system in Canadian youth. *Journal of Sociolinguistics*, 8(4):493–514.

Tagliamonte, S. and D'Arcy, A. (2007). Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change*, 19(02):199–217.

Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2):135–178.

Tannen, D. (1984). *Conversational Style*. Oxford University Press, Oxford.

Tannen, D. (2000). "Don't just sit there–Interrupt!" Pacing and pausing in conversational style. *American Speech*, 75(4):393–395.

Tatman, R. and Kasten, C. (2017). Effects of Talker Dialect , Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Proceedings of Interspeech 2017*, pages 934–938.

Thomson, R. (2006). The effect of topic of discussion on gendered language in Computer-Mediated Communication discussion. *Journal of Language and Social Psychology*, 25(2):167–178.

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambdrige University Press, Cambridge.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Wagner, S. E., Mason, A., Nesbitt, M., Pevan, E., and Savage, M. (2016). Reversal and re-organization of the Northern Cities Shift in Michigan. *University of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 44*, 22(2):171–179.

Wales, K. (2004). Second person pronouns in contemporary English: the end of a story or just the beginning? *Franco British Studies*, 33-34:172–185.

Ward, M. (2003). *Portland Dialect Study: The fronting of /ow,o,uw/ in Portland, Oregon*. PhD thesis, Portland State University.

Weninger, F., Marchi, E., and Schuller, B. (2012). Improving recognition of speaker states and traits by cumulative evidence: Intoxication, sleepiness, age and gender. In *INTERSPEECH*, pages 1159–1162.

Wieling, M. and Nerbonne, J. (2010). Hierarchical spectral partitioning of bipartite

graphs to cluster dialects and identify distinguishing features. In *ACL 2010*, pages 33–41.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Systems, Man, and Cybernetics*, 2:408–421.

Wolfram, W. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Center for Applied Linguistics, Washington, D.C.

Wolfram, W. and Schilling, N. (2015). *American English: dialects and variation*. John Wiley & Sons.

Wong, A. (2007). Two vernacular features in the english of four American-born Chinese. *University of Pennsylvania Working Papers in Linguistics*, 13(2):217–230.

Wong, A. W.-m. and Hall-Lew, L. (2014). Regional variability and ethnic identity: Chinese Americans in New York City and San Francisco. *Language & Communication*, 35:27–42.

Xiao, R. and Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies*, 1(2):241–273.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2017). Achieving Human Parity in Conversational Speech Recognition. Technical Report Microsoft Research, Microsoft.

Xue, S. A., Hao, G. J. P., and Mayo, R. (2006). Volumetric measurements of vocal tracts for male speakers from different races. *Clinical Linguistics & Phonetics*, 20(9):691–702.

Yaeger-Dror, M. and Thomas, E. R. (2010). *African American English Speakers and their Participation in Local Sound Changes: A comparative study*. Duke University Press for the American Dialect Society, Durham.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Yuasa, I. P. (2008). *Culture and Gender of Voice Pitch: A Sociophonetic Comparison of the Japanese and Americans*. Equinox, London.