Modeling the Perceptual Learning of Novel Dialect Features

Rachael Tatman

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Richard Wright, Chair

Alicia Beckford Wassink

Gina-Anne Levow

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

Modeling the Perceptual Learning of Novel Dialect Features

Rachael Tatman

Chair of the Supervisory Committee:
Professor Richard Wright
Linguistics

All language use reflects the user's social identity in systematic ways. While humans can easily adapt to this sociolinguistic variation, automatic speech recognition (ASR) systems continue to struggle with it. This dissertation makes three main contributions. The first is to provide evidence that modern state-of-the-art commercial ASR systems continue to preform reliably worse on talkers from some social backgrounds. The second contribution is expanding our understanding of how and when human listeners who have been recently exposed to a new dialect rely more on social information about a talker than the acoustics. While human listeners' perceptions can be categorically shifted by giving them incorrect social information when listening to a new dialect, the same effect is much weaker when listening to their own dialect. The third contribution is computationally modeling listeners' bias towards their own dialect. Models trained using a dataset biased towards one dialect accurately reflected the behavior of listeners from that dialect. Further, explicitly including the dialect from which each training token was drawn during training and providing it at the time of classification improved classification accuracy with the second dialect while maintaining accuracy for the first. This can provide a behaviorally-accountable model for dialect adaptation in automatic speech recognition.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

It has been a true privilege to have the time and resources to devote to writing this dissertation. While much of the labor has been mine, I have been assisted in innumerable ways by colleagues, friends and family and I would like to express my gratitude for this.

First, I would like to thank my committee members: Dr. Alicia Beckford Wassink, Dr. Gina-Anne Levow and, most importantly, my advisor Dr. Richard Wright. Their guidance, knowledge and attention to detail have polished a collection of loosely related ideas into a cohesive research program firmly rooted in existing scholarship. Secondly, I would like to thank the scholars whose previous work built the foundation for my research. In particular, I wish to thank Dr. Catherine Inez Watson for providing the New Zealand English recordings used throughout my dissertation. I would also like to thank the R package maintainers and contributers (including Dr. Daniel McCloy) who have saved me countless hours of implementation.

I would also like to thank the reviewers who have reviewed parts of this dissertation. I have had the privilege of presenting work from this dissertation at the Workshop on Ethics in Natural Language Processing (2017), the Cognitive Modeling and Computational Linguistics Workshop at EACL (2017), the 4th Pacific Northwest Regional NLP Workshop (2016), the 3rd Conference on Experimental Approaches to Perception and Production of Language Variation (2016), the Meeting of the Acoustical Society of America (2016) and the Doctoral Consortium at the EMNLP Workshop on Natural Language Processing and Computational Social Science (2016). The comments and suggestions from reviewers and participants at

these conferences have substantially improved this work and I am deeply grateful for them.

I have also benefited from being a member of an active community of scholars, in both physical and electronic spaces. I have been very fortunate to study in a warm, supportive department, and I want to extend my sincerest gratitude and fondness to the University of Washington linguistics department faculty (especially my committee and Dr. Emily Bender), staff (especially Joyce Parvi, Michael Furr, Catherine Carrera and Dr. Elizabeth McCullough) and current and former students (especially Kirby Conrad, Dr. Valerie Freeman, Taylor Carrasco-Hermerding, Amandalynne Paullada, Esther Le Grezause, Dr. Sarala Puthuval, Stephanie Peterson, Laurie Dermer, Leanne E. Rolston, Rik Koncel-Kedziorski, Katy King and Dr. Hyunjung Sophie Ahn, among many others). You have collectively created and maintained an environment of convivial scholarship and I will genuinely miss working alongside you all. In addition, I have immeasurably benefited from connecting with colleagues on-line, especially on Twitter. I would particularly like to acknowledge Gretchen McCulloch (@GretchenAMcC) for her central role in fostering the internet (not Internet ;) linguistics community.

I think it important to acknowledge the financial assistance I have received, especially the National Science Foundation Graduate Research Fellowship (grant number DGE-1256082). This allowed me to support myself without additional employment during the last three years of my five-year doctoral program, and has contributed in no small way to my timely completion.

Last, but far from least, I would like to extend my gratitude to my family, including my father, my mother (whose collegiate studies have closely paralleled mine and who graduated from the University of Virginia with a bachelor's in sociology mere weeks before I finished my doctorate), and my sibling. And I would be beyond remiss if I did not acknowledge the

loving support of my husband, Conner Kasten. His companionship, emotional and material support and last-minute copy-editing have helped me not just finish my dissertation, but also maintain a high quality of life while doing so.

Chapter 1
INTRODUCTION

Consumer automatic speech recognition (ASR) technology is increasingly popular. One 2014 Google study estimated that in the United States 55% of teens and 41% of adults used voice search at least once a day (Huffman, 2014). And the pace of adoption has not slowed: a 2016 survey found that only 2% of iPhone users and 4% of Android users did not use each system's built-in voice assistants (Milanesi, 2016).

Despite their increasing reach, automatic speech recognition differs from human speech recognition in ways that make it less usable than it could be. In particular, while accounting for sociolinguistic variation is an intrinsic part of human speech perception, sociolinguistic variation, including dialect differences, continues to be a hurdle for automatic speech recognition (Benzeghiba et al., 2007).

Automatic speech recognition has the capacity to improve the human experience. It can be particularly beneficial to individuals who may otherwise struggle with access to technology, whether due to illiteracy, disability or situational factors by allowing them to interact with computers without relying on text or touch. But in order to best serve all users it should be comparable in its ability to adapt to variation to the system it is trying to mimic: human speech perception.

1.1 Social variation is intrinsic to speech

Many aspects of individuals' identities are reflected in their language, including age, gender, social class, ethnicity and regional origin (Chambers, 1995). Quantifying this is the focus

of variationist sociolinguistics, a field which began with Labov's investigation of Martha's Vineyard, originally published in 1963 (Labov, 1963). Since then, our understanding of sociolinguistic variation has expanded significantly to include the influence of individuals' peer groups (Milroy, 1980; Eckert, 1989a) and stylistic choices (Bell, 1984; Giles, Coupland, & Coupland, 1991). While this remains an active area of research and there is undoubtedly much that we do not yet know, we can confidently state that all language contains predictable variation that arises from the social identities and stylistic choices of talkers.

Human language understanding, and particularly speech perception, does not occur *in spite* of this variation but rather in concert with it; speech perception is inextricably tied to the listener's social perception and evaluation of the talker. This social perception does not occur after listening to speech but during processing. There is evidence that listeners notice that a new talker has a different dialect within 200ms–before they finish the first syllable, let alone the first word (Scharinger, Monahan, & Idsardi, 2011). And not all social information comes from the speech signal alone. Listeners also rely on extra-linguistic information during speech perception. Cues as subtle as a stuffed animal associated with a particular language variety (like a kangaroo for Australia and a kiwi bird for New Zealand) can shift a listener's perceptions towards that variety (Hay & Drager, 2010). These differences between language varieties, while initially confusing, can be learned. After sufficient exposure any listener can learn to easily comprehend multiple dialects, even if there are strong differences between how words are produced in these dialects (Sumner & Samuel, 2009; Baese-Berk, Bradlow, & Wright, 2013).

While a great deal is known about the human capacity to learn and understand speech across language varieties, there is still more to discover, especially around the role of extra-linguistic social knowledge of the talker. This dissertation focuses on one aspect of this problem: once a listener has learned to recognize the sounds of a dialect that is new to them,

how much do they rely on extra-linguistic knowledge about a talker (like what country they're from) as opposed to information in the speech signal itself? Secondly, does the use of extra-linguistic and acoustic information change depending on a listener's level of experience with a dialect?

The answers to these questions are of particular relevance to work on accent adaptation for automatic speech recognition. There is increasing interest in using extra-linguistic information, such as GPS data, during speech recognition (Chelba, Zhang, & Hall, 2015, for instance). Explicitly computationally modeling how human listeners are using both extra-linguistic and acoustic information to recognize speech in potentially multi-dialectal situations can provide a roadmap for ASR systems doing the same task.

1.2 Should computational linguistic interfaces be human-like?

This dissertation is built on a foundational assertion that automatic speech recognition, and in particular adaptation to new dialects, should be human-like. There are certainly areas where it is useful for an ASR system to diverge from human speech perception, such as recognizing speech at a much faster rate than a human would. However, especially for voice assistants, the primary purpose is for a user to interact with the system as if it were a human listener. Any conversational agent will be unavoidably compared to users' other conversational partners: humans. Systems that consistently make perception errors that a human listener would not are both frustrating and less useful than they could be.

One area where ASR errors are particularly salient is when voice commands are the users' primary way of interacting with a system. While voice-controlled virtual assistants are a relatively recent phenomena, people with disabilities have been using voice interfaces for much longer. This is an area where I have had personal experience. I am dyslexic, and as a child struggled with writing. In the fourth grade, the school I was attending had me

try to use dictation software for course assignments. However, the sheer number of errors the system made meant that it took me far longer to dictate and correct the resulting errors than just to (with great difficulty) type in the first place. I vividly remember dictating "the walls were dark and clammy" and having it transcribed as "the wells were gathered and planning". Granted, children's voices are difficult for speech recognition systems trained primarily on adult speech (which most are) and overall accuracy has improved dramatically in last seventeen years. However, the frustration of this experience, where voice technology that was intended to make writing easier instead made it more difficult, has remained vivid and colors my experience with technology to this day. And low accuracy is even more frustrating when speech recognition is used, not only for dictation, but for command and control.

But how can ASR become more human like? One way is to apply the knowledge that linguists have about how human speech perception works. This is not a new stance: many scholars working on both human speech recognition and automatic speech recognition have pointed out that applying insights from human studies has the potential to dramatically improve ASR (Scharenborg, 2005; Baker, Eddington, & Nay, 2009, among others).

One key area where humans are clearly superior is perception in the face of social variation. For a human listener, variation in the speech signal is not a hindrance to comprehension, but rather an integral part of speech perception. Automatic systems, on the other hand, often struggle to maintain accuracy across populations of talkers, especially if the acoustic training data comes from a relatively socially homogeneous population of talkers. This arises from the convergence of two processes. The first is the sociolinguistic variation discussed previously: we know that socially-meaningful groups of language users use language in similar ways. The second is the fact that acoustic models are trained by reducing error on their training dataset. In other words, speech recognition systems are most accurate on speech

that is more like that it was trained on. And if the talkers who are using it speak in a way that's different from the speech it was trained on due to sociolinguistic variation, those talkers will have a higher error rate.

1.3 Consequences of linguistic discrimination in ASR systems

Why is it so vital that ASR systems perform equally well for talkers with a range of social backgrounds? Higher error rates in are frustrating for users, of course, and may result in them spending marginally more time correcting those errors, but is this really a pressing problem? Granted, the consequences of higher error rates in consumer ASR systems are not dire. However, the stakes are raised considerably when these systems are incorporated into products which directly affect user's and other stakeholder's quality of life. If errors disproportionately affect one social group more than another, this is a form of linguistic discrimination.

The first instance where higher error rates have a direct, negative impact on their users is when the error rate of ASR program directly affects a users' ability to do their job. Medical transcription, where a medical specialist's verbal report is converted into text, was previously done primarily by hand. However, the use of automatic speech recognition in this context has become increasingly common. This is concerning, given work which found that least one medical transcription software program was significantly less accurate for women (Ali et al., 2007). Since errors in medical transcription are high stakes (a similar but incorrect word in a diagnosis can literally be a matter of life and death) careful correction of ASR output is necessary. If one group of users (in this case women) must spend more time than their colleagues correcting errors, it represents a direct reduction of their ability to spend time on other tasks necessary for their job.

Another way in which ASR systems have the potential to negatively affect users is when

the output of ASR systems is incorporated into hiring and promotion decisions. This is not a hypothetical use-case, but one which is becoming increasingly common (Shahani, 2015; L. Morrison, 2017). Based on biases observed in other systems, these applications may be making more errors on speech from women, people of color and talkers who do not speak General American English.

A final instance where differential ASR outcomes have the possibility of directly negatively affecting users is when these systems are used in student evaluation. While I am not currently aware of any state-mandated student testing being conducted using ASR, multiple systems are already in development for this purpose (Yilmaz, Pelemans, et al., 2014; Proença, Celorico, Lopes, Candeias, & Perdigão, 2016). This is concerning since student's test scores are often directly linked to their future potential opportunities: only students with test scores above a certain threshold might be allowed to participant in honors programs, for example. It is possible that such a testing system might not correctly recognize speech from Black students with the same degree of accuracy as others students, especially given research which suggests current NLP are less accurate for African American English talkers (Blodgett, Green, & O'Connor, 2016). If this resulted in Black children receiving lower test scores not because their answers were incorrect, but because the system made an error, this would result in a tangible harm to these students in the form of reduced educational opportunities.

As speech recognition is incorporated into systems applied to a wider range of applications, the potential harm of linguistic discrimination–unintentional though it may be–increases dramatically. Even worse, biases against language forms associated with certain groups of language users have the potential to affect far more than just speech technologies. There is a growing body of evidence that sociolinguistic variation at every level of the grammar–from the phone level, to the lexicon, to the use of specific syntactic structures–is encoded as ro-

bustly in text as it is in speech (Grieve, 2016; Tatman, 2016; Eisenstein, 2017; D.-P. Nguyen, 2017). This has already been shown to result in some language users, in particular Black Twitter users, being under-served by text processing tools (Blodgett et al., 2016). In order to ensure that systems which automatically process language are fair and equitable, we must work to identify and actively mitigate linguistic discrimination.

1.4 How many people might be discriminated against by ASR systems trained on only Mainstream US English?

Approximately how many talkers of Mainstream US English (MUSE), also called "General American" or "Standard(ized) American English" are there in the United States? This is a difficult question, especially given that there is no wide consensus about what MUSE is. There is, however, agreement about what it is not. In particular, educated talkers from the Midwest and West are generally considered to be MUSE talkers by non-linguists (Lippi-Green, 1997, p. 60). I would also argue that non-linguists often characterize talkers of both African American English and Chicano English to be non-MUSE talkers, as well as talkers who are not native English talkers.

Since the United States census asks census-takers about their language background, race and ethnicity, educational attainment and geographic location, we can use census data to roughly estimate how many talkers of MUSE English there are in the United States. I chose to use the 2011 census, as detailed data on language use has been released for that year on a state-by-state basis (Ryan, 2013). From this data, I calculated how many individuals were living in states assigned by the U.S. Census Bureau to either the West or Midwest, and for these states how many surveyed spoke English 'very well' or better. Then, assuming that residents of these states had educational attainment rates representative of national averages, I estimated how many college educated (a bachelor's degree or above) non-Black and non-Hispanic talkers lived in these areas.

Talkers in the 2011 census who are...	Count	Percentage of US Population
Living in the United States...	311.7 million	100%
...and live in the Midwest or West...	139,968,791	44.9%
...and speak English at least 'very well'...	127,937,178	41%
...and are college educated[1]...	38,381,153	12.31%
...and are not Black or Hispanic[2].	33,391,603	10.7%

Table 1.1: Table estimating the number of Mainstream US English talkers in the United States, based on information from the 2011 US census. Note that though this variety is sometimes considered the "default", the majority of talkers do not use it.

As can be seen in Table 1.1, based on the criteria laid out above only around a tenth of the US population are likely to be considered talkers of Mainstream US English by a non-linguist. This estimate is possibly somewhat conservative. In particular, not all Black talkers use African American English and not all Hispanic talkers use Chicano English, and the regional dialects of some parts of the Northeast are also compatible with the popular perception of Mainstream US English. However, even if this estimate were wildly conservative and a full 50% of the United States population can be considered talkers of some type of "standardized" dialect, any tools which only or preferentially serve this population will be ignoring the reality of linguistic diversity.

1.5 Contents of this dissertation

In addition to the introduction (this chapter), this dissertation has six chapters which present a combination of system evaluations, experimental work and computational modeling and a conclusion chapter.

Chapter two is an in-depth discussion of previous research relevant to this work. Since the research questions posed here span multiple disciplines, part of the contribution of this work is drawing together related research from different fields. In particular, the fields of perceptual learning, linguistic social priming and work on the task of accent adaptation in

automatic speech recognition are discussed.

The third chapter documents that, despite major advances in overall accuracy, social biases are encoded in current state-of-the-art speech recognition systems: Google's speech recognition (as implemented in YouTube's automatic captions) and Microsoft's Bing Speech API. I have found strong evidence of differences in accuracy across dialects, as well as evidence for differences in accuracy by talker gender and ethnicity.

The fourth chapter presents experimental work on the role of social information in speech perception after dialect learning. Listeners from the United States who had successfully learned to identify the vowels of a talker of New Zealand English where able to transfer this learning to a second talker when they were told that they were listening to a New Zealand English talker. However, when incorrectly told they were hearing a talker from the United States, they "undid" their recent learning and identified vowels as if the talker really were from the US.

In chapter five, listeners' use of social information is modeled using conditional inference trees, a type of unbiased decision tree that splits based on statistical inference. When a talker's dialect was included as a feature, it resulted in a much more human-like error pattern. In particular, when the model was given incorrect labels, it produced the same types of errors observed in the experiment.

Chapter six replicated the experiment presented in chapter four, but with a fully crossed design. After learning, listeners were presented with both a new New Zealand English talker as well as an American English talker. When told the correct national origin of each, listeners were largely accurate in identifying vowels. However, when they were mislabeled, listeners only dramatically shifted their classifications for the New Zealand English talkers. When listening to their own dialect, however, listeners were able to gain enough social information from 300ms worth of acoustic information to ignore the incorrect label. Regardless of the

extra-linguistic information they were given, listeners correctly identified the vowels of their own dialect.

The seventh chapter discusses the creation of a computational model capable of replicating listener's judgment patterns. As in chapter five, listener classifications were modeled with conditional inference trees. Bias towards correct classification of one dialect was captured by biasing the data used to train the classifier. A classified trained with very biased data (90% drawn from one dialect) accurately captured listener's perceptual bias towards their own dialect.

Finally, chapter eight is a conclusion which draws together the rest of the dissertation. It includes a discussion of the work presented here, its implications for both linguistics and automatic speech recognition and directions for future work.

Chapter 2
BACKGROUND

This dissertation builds on three main areas of research. The first is perceptual learning, the long-term re-organization of the perceptual system that allows an organism to better interact with its environment (Goldstone, 1998). It is a framework that has been fruitfully applied to a number of areas of speech perception (Samuel & Kraljic, 2009). The specific focus of this project is the perceptual adaptation that occurs when listeners first encounter a new and unfamiliar variety of a language they already know.

Second, this dissertation builds on prior work in linguistic social priming, specifically the body of work in sociolinguistics and speech perception. This is where a listener's beliefs about a talker are shown to change their percepts. For example, a talker from Alabama who has the pin-pen merger might say "pen" [pɪn]. However, if that same audio token were presented as if said by a talker from Washington State who does not have the pin-pen merger, a listener might instead report that they heard the word "pin".

The third area of research relevant to this dissertation is automatic accent adaptation. This is the computational task of adapting an existing ASR system to better perform on a talker from a new dialect. While this dissertation focuses on modeling behavioral data, rather than implementing a novel accent adaptation algorithm, I will discuss the approaches that have already been developed and compare their performance with that of humans doing a parallel task in order to provide a road-map to implementing more human-like accent adaptation.

2.1 Perceptual Learning

2.1.1 How is perceptual learning relevant to this work?

Talkers of a language can quickly adapt their perceptions in order to understand talkers of a dialect that differs from their own, even if they have never encountered this dialect before. This process is often referred to as "dialect adaptation" and falls under the more general heading of perceptual learning.

2.1.2 Perceptual learning

The pioneer of perceptual learning research was Elanor Gibson (not to be confused with her husband, James Gibson, who also researched perception) who launched the field of inquiry with her 1969 book. Gibson rejected earlier theories that perception itself had to be learned. These held that information is initially taken in in an unstructured way and compiling seemingly unrelated information into a cohesive percept (such as "ball" or "tree") requires learning. Under this earlier view infants would need to learn how to perceive "depth" by compiling different types of visual information, such as areas of contrasting light and dark. However, Gibson successfully showed that infants do have robust–and, she argued, innate–depth perception using the now well-known "visual cliff" experiment (Gibson & Walk, 1960). In it, infants could not be made to voluntarily crawl onto a plane of glass which attached to the edge of a table. While perfectly safe, the infants perceived that there was a steep drop and refused to cross it. So in its earliest iterations, "perceptual learning" was the argument that, rather than having to learn how to perceive, humans use perception in order to learn. The natural extension of this was that learning can then result in better use of perceptual information (Gibson, 1969). Perhaps the best known example of this type of learning is the ability to better distinguish between two very similar things, or differentiation.

Some examples of differentiation include the ability to tell whether a glass of wine is from the top half or bottom half of the bottle, or distinguishing the sex of newly-hatched chicks (Gibson & Walk, 1960)

Despite the amount of interest in perceptual learning in the 1960's, this line of inquiry then lay largely dormant for several decades (Kellman & Massey, 2013). Perceptual learning, and particularly differentiation, enjoyed a renaissance in the 1990's and was used as a framework to investigate speech perception. At the forefront of this renewed interest was Goldstone's 1998 review paper (Goldstone, 1998), which outlined four central types of learning which fall under the purview of "perceptual learning". The one which is most applicable here is differentiation, which he proposes can be divided into four different types: differentiation of whole stimuli (which includes psychophysical differentiation), differentiation of complex stimuli, differentiation of categories and differentiation of different dimensions.

Differentiation of whole stimuli is an important part of speech perception. Goldstone describes three different facets of this task. The first, psychophysical differentiation, covers a variety of ways in which perceivers psychologically warp the objects of perception in order to aid differentiation. While this is an area of ongoing research (Houtgast, 1995; Jiang & Bernstein, 2011, for example), it is outside the scope this dissertation.

The second type, and most canonical example of perceptual learning for speech, is that of complex stimuli. Goldstone uses the example of training Japanese talkers learning English to distinguish /l/ from /r/, a distinction which does not exist in their native language (Lively, Logan, & Pisoni, 1993). Especially for researchers in second language acquisition, this has proven a rich seam of research (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Best, McRoberts, & Goodell, 2001; Bradlow & Bent, 2008).

The final type of whole stimuli differentiation deals with learning to identify sub-sets of categories–for example listeners who already have the ability to identify classes of speech

sounds learning to differentiate between nasalized and non-nasalized vowels (Harnsberger, 2001) or rounded and unrounded vowels (Meers, 2009).

By differentiation of dimensions, Goldstone means the ability to break stimuli down into their component parts. For instance, while children may be able to tell that there's a difference between very similar toys, they may struggle with identifying which dimension they differ in, like size or color. A similar example from speech perception is the ability to correctly distinguish that a sequence of two tones is different, without the ability to determine whether they're rising or falling. While Mandarin speakers are able to correctly identify lexical tone with much higher precision than English speakers, they actually fared worse on a two-note contour identification task: English talkers reached ceiling with a difference between tones of only 30 Hz, while Mandarin talkers needed a 40 Hz difference (Bent, Bradlow, & Wright, 2006). This suggests that while they have very good ability to differentiate complex stimuli, the Mandarin-speaking participants may have less accurate differentiation in just the pitch dimension.

Perceptual learning, and particularly differentiation, is a framework that can be fruitfully applied to speech perception, and it has been in the almost two decades since perceptual learning was revived as an active research area. A key question of this dissertation is how listener's social identity affects both perceptual learning, and perception after learning.

2.1.3 Talker-specific learning

It is well established that listeners can learn to adapt to specific talkers, an effect which is often discussed within an episodic encoding or exemplar based framework (Johnson, 1997). Early research showed an effect described as the "familiar talker advantage". This is the propensity for listeners to more easily and accurately perceive the speech of talkers that they had previously been exposed to (Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni,

1998). There's also evidence that talker identity is an innate part of language learning–that is, we don't learn how a word is said so much as we learn how a particular talker says that word (Creel & Bregman, 2011).

In some studies, dialect learning has been attributed to talker specific effects. Trude and Brown-Schmidt, for example, found that listeners were able to simultaneously shift their perceptual boundaries for one talker while leaving them unchanged for the second (Trude & Brown-Schmidt, 2012). This effect was intensified when information about which talker the participant was hearing was presented. It is possible, however, that this effect may be the result of listeners determining that these talkers did not share a dialect.

In addition, talker-specific learning cannot account for observations made about dialect learning, like the fact that being exposed to a variety of talkers of a novel variety assists learners more than just being exposed to one. Research on training Japanese listeners to identify English /r/ and /l/ found that training data which included a large amount of variation led to improvements in listener's ability to discriminate between these sounds (Logan, Lively, & Pisoni, 1991). And Bradlow and Bent found that English speakers could more easily perceptually adapt to novel speakers of Chinese-accented English if they had been previously exposed to multiple different Chinese-accented talkers (Bradlow & Bent, 2008). This suggests that, in addition to talker-specific learning, listeners use information about groups of talkers.

2.1.4 Perception and social categorization

"Social categorization" is a term from social psychology and refers to the way in which individuals perceive themselves and others as members of social groups and the effects of these perceptions on cognition (Turner & Oakes, 1986). There is a large body of evidence that listeners are accessing their beliefs about groups of talkers during speech perception in addition

to prior knowledge of specific talkers. Listeners use information on talkers' social characteristics–including gender, age and likely regional origin of language background–during speech perception.

Gender: Early work on the perceptual effect of listener's beliefs about a talker's gender found that the gender of a face shown along with the same audio shifted the boundary between the phonemes [ʊ] and [ʌ] (Johnson, Strand, & D'Imperio, 1999). In her dissertation, Strand found that talkers were generally slower in a speeded repetition task when presented with a face which did not match the gender of the voice they heard (Strand, 2000).

Age: Listeners' beliefs' about talkers' age has also been shown to affect perception. For instance, eye-tracking data has shown that listeners from Texas are more likely to expect to hear a merged pin/pen when shown a picture of an older talker (Koops, Gentry, & Pantos, 2008). There is also evidence that listeners are both faster and more accurate in identifying a word used more often by talkers from a certain age cohort when it is said by a congruently-aged voice (Walker & Hay, 2011).

Regional origin, language background and dialect: The body of work most relevant to this dissertation, however, looks at the effect of listener's beliefs about a talker's regional origin and language background on their speech perception. One of the foundational studies in this research area, by Nancy Niedzielski (Niedzielski, 1999), investigated how listeners' integrated their social information into their perception. Listeners from Detroit were played a set of resynthesized vowel tokens and asked to pick which one best matched the speech of a speaker from either Detroit or Canada. When asked to pick which vowel was most like one produced by a speaker from Detroit, listeners picked a vowel which was Northern Cities Chain-Shifted, while for perceived Canadians they picked one which was not.

This, Niedzielski argued, was evidence that listeners use social information in speech perception. Further work supported this claim. Sidaras and co-authors found that listeners were able to quickly adapt to Spanish-accented English and apply this learning to new speakers (Sidaras, Alexander, & Nygaard, 2009), results which were very similar to the work on Chinese-accented talkers discussed above (Bradlow & Bent, 2008). There is also evidence that listeners are more adept at transcribing Chinese-accented speech when presented with a photo of a Chinese individual (McGowan, 2015). Even very subtle cues to the perceived social identity of a talker, such as the presence of a stuffed kangaroo or kiwi, can affect perception to be more in line with Australian or New Zealand English respectively (Hay & Drager, 2010). Further evidence for listeners' use of their social expectations comes from Szakay et al., who found the expectation of a talker's Māori identity results in cross-linguistic priming from Māori to Māori English (Szakay, Babel, & King, 2016). There also seems to be some benefit to intentionally training listeners in social categorization: when listeners were presented with talkers who they had previously been exposed to, explicit cues about the talker's identity boosted perception (Trude & Brown-Schmidt, 2012).

Learning new dialects: This dissertation focuses on cases where the social groups of interest are talkers who use the same dialect. This process of dialect adaptation is of interest as a research question both because it closely parallels the computational task of accent adaptation, and also because there is evidence that learning and perceiving different varieties of the same language is a unique task, even at the neurological level (Floccia, Goslin, Girard, & Konopczynski, 2006; Scharinger et al., 2011; Adank, Nuttall, Banks, & Kennedy-Higgins, 2015).

One of the earliest studies that looked at the process of perceptually learning regional dialects found that playing listeners two-minute examples of regional speech pushed catego-

rization of subsequent vowel productions towards those that the listeners had experienced earlier (Evans & Iverson, 2004). (This was in contrast to earlier work on perception of foreign-accented speech, where listeners maintained their own distinctions and perceptually warped, or "corrected", talkers' production (Best et al., 2001).) Critically, these shifts occurred very quickly and persisted for the duration of the experiment. Later work by the same researchers, however, failed to find that listeners had shifted their perceptions, even with long-term exposure to a different regional dialect (Evans & Iverson, 2007). The authors suggested that this may be because dialect perception occurs on a talker-by-talker basis. It is also possible, however, that listeners were affected by other factors, such as their stance towards the second dialect. Nguyen and co-authors, for example, find a complex interplay between listeners' attitudes towards a dialect, prior experience with it and their expectation of which dialect they believed a talker would use (N. Nguyen, Shaw, Pinkus, & Best, 2016). There is also evidence that listeners adapt not only to specific accents but also the expectation that there is variation in accents. Baese-Berk and co-authors found that listeners who were exposed to talkers from five language backgrounds were able to generalize to talkers from language backgrounds they had not heard during training (Baese-Berk et al., 2013).

There is also a body of work that suggests that listeners' own dialect affects their perception. For example, participants in a speeded word classification task performed the best when their own dialect was presented in isolation. Listeners' perception was impaired on both a different dialect or a mix of their own and a different dialect (Clopper, 2007). Further, vowel identification rates for listeners from Kansas have been found to vary significantly across other American dialectal regions (Oder, Clopper, & Ferguson, 2013). In addition, which cues a listener uses during adaptation depends on their dialect. Karpinska and co-authors found that, for listeners from the UK, New Zealand, Ireland and Singapore, formants were the most important cues distinguishing the vowels in "bit" and "beat". For Australian listen-

ers, however, duration was a more important cue to the identity of these vowels (Karpinska, Uchida, & Grenon, 2015).

However, it is possible for a listener who does not use a dialect to still perceive it with ease. Sumner and Samuels investigated the perception of listeners from one American English dialect area who had moved to another (Sumner & Samuel, 2009). They found evidence that listeners who had extensive experience with one dialect but were not speakers of it showed the same cross-dialectal priming patterns (e.g. a word from New York English priming the same word in General American English and *vice versa*) as speakers of that dialect.

Considering listeners' social categorization of talkers may help to explain results that are otherwise surprising. For instance, Eisner and McQueen found that when they were presented with new voices, listeners who had previously shown evidence of perceptual learning lost the effect (Eisner & McQueen, 2005). However, this may be due in part to the fact that there was no reason for listeners to assume that novel talkers belonged to the same social grouping as the training talkers–especially since the training talkers had a salient "dialect" feature.

A second study looked at whether listeners would show a benefit of training on a voice speaking one language if that talker was speaking a different language (Levi, Winters, & Pisoni, 2011). In this study, listeners trained on talkers speaking English showed improved accuracy when listening to those same talkers saying novel English words. However, listeners trained on talkers speaking German showed no benefit when listening to those voices speaking novel English words. Crucially, the authors did not report that listeners–even those who achieved high voice recognition in the training session–were made aware that they were listening to the same talkers. Thus, it is entirely possible that listeners, hearing tokens in a different language, socially categorized the voices in the word-recognition test as distinct from those they had been trained on.

The idea that a listener's social expectation affects their speech perception is not novel.

Previous researchers have explicitly called for perceptual models to include social information, both for general cross-modal perception (Xiao, Coppin, & Van Bavel, 2016) and speech perception in particular (Foulkes, 2010). Babel and Russell (2015) show that there was a processing cost for listeners when their social expectations didn't align with social information about the talkers (in their case images of people from different ethnic groups along with matched and mismatched audio), suggesting a central role of the use of social information during speech perception . Kleinschmidt and Jager state that "...good speech perception depends on using an appropriate generative model for the current talker, register, dialect, etc." (Kleinschmidt & Jaeger, 2015). And Dahan and collaborators suggest a similar mechanism might play a role in perceptual adaptation: "listeners may initially assume that their experience exemplifies something very general about the talker (or even a speech community) and display broad generalization" (Dahan, Drucker, & Scarborough, 2008). Further, we know that listeners don't use all the social information available to them, but rather are more likely to rely on robust social categories such as gender (Rácz, Hay, & Pierrehumbert, 2017; Samara, Smith, Brown, & Wonnacott, 2017), and that the ability to use social categorization during speech perception is acquired over time, with older children better able to do so than younger children (Levi, 2014). Taken together, these findings suggest that it would be wise to explicitly incorporate ideas of how listeners are engaging in social categorization during perceptual learning.

2.2 Automating dialect adaptation

Since a significant component of this work will involve computational modeling of perception and perceptual adaptation, it makes sense to cover related work. Perception of different dialects, and especially adaptation to new dialects, is an area that has received a good deal of attention in the computational literature, especially in the area of Automatic Speech

Recognition (ASR).

This is not intended to a be a through review of all automatic speech recognition systems designed to deal with accent variation (since this is not an ASR dissertation), but rather a sampling to help establish common methods in the field. The majority of work makes use of two main implementation strategies: the first is adaptation to individual talkers and the second is an attempt to automatically detect a talker's dialect and then apply a recognizer designed for use with that dialect. While neither of these methods of accent adaptation are explicitly meant to be models of perceptual learning, or human behavior at all, it is worth considering how they match up with our knowledge of how humans accomplish the same task. It is possible that basing a model of talker adaptation on behavioral models of perceptual learning would result in a less desirable performance from an ASR point of view, such as computational inefficiency or human-like errors. It is still worth building such models, however. In particular, behavioral scientists can benefit from an additional line of evidence for theoretical ideas and, if human-like behavior is desired for certain applications, it will be useful to have an explicit quantitative model of perceptual adaptation.

First, however, a few notes on terminology: the work discussed here is drawn mainly from the fields of natural language processing and speech processing, and the use of terminology differs between these fields and linguistics. In particular, the term "accent" is often used where a linguist might prefer the term "dialect". "Accent" has a broader meaning than dialect, however, as it is also often applied to speech produced by non-native talkers, while "dialect" rarely is.

2.2.1 Talker adaptive models

Talker adaptive models are models which change in order to achieve better performance on specific talkers. They may do this iteratively over time, or all at once at an early stage

of model construction. Some of the more popular techniques in talker adaptation include Maximum A Priori (MAP) adaptation, Maximum Likelihood Linear Regression (MLLR) and eigenvoices. These generally build on standard speech recognition models trained using Hidden Markov Models (HMM) and with their output probabilities expressed as Gaussian Mixture Models (GMM) (Rabiner & Juang, 1993).

Maximum A Priori (MAP) adaptation, or MAP, is a form of adaptation where the input model parameters are used as the prior and the talker-adapted classification is the most likely classification given that prior and training data from the talker (Gauvain & Lee, 1994). There are several drawbacks to this method. The first is that it requires a large quantity of data for successful adaptation, and that even with very large training data sets the number of parameters in a model is so large that it is likely that most will not be updated during training.

Maximum Likelihood Linear Regression, or MLLR, is a talker adaptation technique that allows individual parameters or groups of model parameters to be shifted independently (Leggetter & Woodland, 1995). This is commonly done, as in the Hidden Markov Model Toolkit (S. Young, Woodland, & Byrne, 1993), with a regression class tree. In a regression class tree, the top node of the tree contains all the parameters, and each node recursively splits the remaining components into two successively more acoustically similar subclasses. The final leaves of the tree (which may, for example, represent individual consonants or vowels) are the base regression classes. In order to adapt to a new talker, then, each node of the tree can be transformed. This may mean updating the model at different levels of the tree, from the top node (which would affect the classification for all speech) to a single base regression class (which might only affect the classification of a single consonant). One benefit of this is that it updates unobserved components. For (an unlikely) example, if no /f/ was observed in the data set but a number of /s/'s were and /s/ and /f/ are grouped

together, then the model will be adapted for both of them. It also requires significantly less training data than MAP adaptation. MLLR remains one of the popular talker adaptation techniques. The main disadvantage is that, since classes of sounds are shifted together, over-adaptation may occur (Ganitkevitch, 2005). In the /s/ and /f/ example, it might be that only /s/ should have been adapted to, but because the parameters for the entire class were transformed, the model /f/ will be erroneously transformed as well.

Eigenvoices (Kuhn, Junqua, Nguyen, & Niedzielski, 2000) are a way of, from a mathematical standpoint, reducing the complexity of speech by reducing the number of dimensions. This is often done by finding the most informative dimensions doing a Principal Components Analysis, or PCA, on the parameters in the input model. The first few dimensions of the PCA analysis (which may not intuitively map to salient perceptual dimensions) are then used to define the eigenspace. In this eigenspace, each voice is represented by a single point, and the distance between voices can be explicitly mapped. Once the position of the voice is determined, this information can be used to transform the original model into one that ideally achieves higher rates of recognition on that specific voice by moving the model towards the voice. Some researchers have combined Eigenvoices with other methods, such as using them as the prior for MAP talker adaptation (Chen & Wang, 2001), or with MLLR transforms (in place of the regression class tree described above) (N. J. Wang, Lee, Seide, & Lee, 2001). While a very flexible and popular technique, eigenvoice talker adaptation does have some drawbacks. The original construction of the eigenspace, though it only needs to be done once, can be quite computationally intensive, and if based on a PCA it requires that all states are represented in both the original model and training data for the new talker, which is not always feasible (Smit, 2010).

Conceptually, adaptation of speech processing systems to individual talkers parallels talker-specific learning and the familiar talker advantage.

2.2.2 ASR and accents

One of the more popular methods in recognizing accented speech is to use accent-specific models. One early approach was to transcribe small amounts of speech from different dialects and then automatically detect phonological differences between accents by creating a decision tree (Humphries, Woodland, & Pearce, 1996). Such accent-specific models continue to be popular and perform well. Of course, this approach requires the ability to automatically detect which accent a talker has, so accent identification has emerged as a related research topic (Lincoln, Cox, & Ringland, 1998; Huang & Hansen, 2007; Aubanel & Nguyen, 2010; William, Sangwan, & Hansen, 2013).

These models, which have as their underlying assumption that groups of talkers talk alike, conceptually parallel the dialect adaptation models discussed above.

Accent models have been shown to out-perform talker adaptation models in some situations. One study compared i-vector and phonotactic accent identification systems to unsupervised talker adaptation as well as a blend of the two methods (Najafian, DeMarco, Cox, & Russell, 2014). The authors found that, while any amount or manner of adaptation improved baseline recognition, accent-identification models out-performed talker adaptation models. Interestingly, accent identification followed by talker adaptation showed the greatest improvement over the baseline.

Another approach is to model individuals as a collection of dialect features (Tjalve & Huckvale, 2005). This has the benefit of allowing for talkers who might have a collection of features from different dialects, as well as being more robust against dialect change. This approach can account for instances where, for example, a younger talker no longer has a dialect feature that older talkers commonly do, but does have features which usually co-occur with that.

2.2.3 Using extra-linguistic data during speech processing

One interesting recent direction in the accent adaptation literature is the incorporation of non-acoustic features to assist in accent adaptation, particularly in accent identification.

One project incorporates information about the geographic location of talkers into the acoustic model directly, and uses a talker's location (taken from their GPS) as a feature during speech recognition (Ye, Liu, & Gong, 2016). Another project actually used no acoustic features for accent identification, instead relying on an automated visual analysis of video footage of talkers (Georgakis, Petridis, & Pantic, 2016).

These projects show that extra-linguistic information can be successfully incorporated into speech processing, which parallels listeners' use of their knowledge about talkers during speech perception.

Chapter 3

IS DIALECT ADAPTATION FOR ASR NECESSARY?

3.1 Evidence of dialect bias from commercial automatic speech recognition systems

In light of the rapid rate of performance improvement in automatic speech recognition (ASR)–a recent Microsoft system achieved a word error rate (WER) of just 6.3% on the Switchboard corpus (Xiong et al., 2016)– it is worth discussing whether accent adaptation still represents a necessary line of research. While deep neural networks and larger training sets have both undoubtedly led to serious improvement in ASR, I will demonstrate here that language variation continues to provide a source of avoidable error. These differences in accuracy are particularly disturbing given that they represent differences in ASR usability for talkers from different social backgrounds in terms of gender, dialect and race.

Though gender-based differences in speech are not the focus of this dissertation, it was included in this investigation for two reasons. First, the existence of gender-based speech variation has been extensively documented in the sociolinguistics literature (Trudgill, 1972; Eckert, 1989b, among many others). Second, gender differences in speech recognition accuracy have been previously reported, with better recognition rates reported for both men (Ali et al., 2007) and women (Goldwater, Jurafsky, & Manning, 2010; Sawalha & Abu Shariah, 2013).

Previous work has also found evidence of dialectal bias in speech recognition, where one dialect was recognized with higher accuracy than another, in both English (Wheatley & Picone, 1991) and Arabic (Droua-Hamdani, Selouani, & Boudraa, 2012). In addition, there

are many anecdotal accounts of bias against dialect in speech recognition. For example, in 2010 Microsoft's Kinect was released and, while it shipped with Spanish speech recognition, it did not recognize Castilian Spanish (Plunkett, 2010).

Finally, there is large body of work showing that there are systematic differences between General American (often associated with white talkers) and varieties such as African American English (Bailey & Thomas, 1998; Rickford, 1999; Cutler, 1999, among many others) and Chicano English (Peñalosa, 1980; Fought, 2002). Given these differences, especially if white talkers are over-represented in the data used to train ASR systems, it is possible that talkers of different ethnic and racial backgrounds may have systematically different error rates.

3.2 YouTube's automatic captions & the accent tag

3.2.0.1 Method

Data for the first half of this chapter was collected by hand-checking YouTube's automatic captions (Harrenstien, 2009) on the word list portion of accent tag (discussed below) videos. YouTube's automatic captions were chosen for three reasons. The first is that they're backed by Google's speech recognition software, which is both very popular and among the more accurate proprietary ASR systems (Liao, McDermott, & Senior, 2013). The second is the fact that accuracy of YouTube's automatic captions specifically are an area of immediate concern to the Deaf community and is a frequent topic of (frustrated) discussion: they are often referred to as "autocraptions" due to their low accuracy and the fact that content creators will often rely on them instead of providing more accurate manual captions (Lockrey, 2015). Finally, YouTube's large user base allowed for the direct comparison of talkers from a range of demographic backgrounds.

Accent tag The accent tag, which was based on Bert Vaux's Harvard dialect survey (Vaux & Golder, 2003), has become a popular and sustained internet phenomenon. In Vaux's own words: "In 2011 or thereabouts, part of my Harvard survey went viral on various social media outlets (youtube [sic] and tumblr [sic] being the most popular), usually under the name 'regional dialect meme', 'accent tag' or 'accent challenge'. As a result, there are countless videos of individuals responding to the survey on the net, creating a mine of information on current English accents - not only from the US, but across the world[1]." In other words, people began recording themselves reading a sociolinguistic word-list and answering lexical elicitation questions and then publicly sharing them online *en masse* for no other reason than because they found it fun and interesting to do so.

Currently the accent tag is perhaps the largest informally-created, publicly-accessible dialect survey ever created. The informal nature of the recordings has both benefits and drawbacks. On the plus side, participant's speech is informal and unscripted, and the viral nature of the accent tag means that there are literally thousands of videos with parallel content uploaded by English speakers from around the world–ideal for sociolinguistic inquiries. The largest drawback, particularly for an acoustic analysis, is the large amount of variation in the recording environments and quality. However, for this initial study, the convenience of being able to directly compare a large number of talkers from different dialect regions reading the same word list outweighed the drawbacks.

Though the Harvard Dialect survey questions, which the accent tag is based on, was designed to elicit differences between dialect regions in the United States, it has achieved wide popularity across the English speaking world. Variously called the "accent tag", "dialect meme", "accent challenge" or "Tumblr/Twitter/Youtube accent challenge", videos in this genre follow the same basic outline. First, talkers introduce themselves and describe their

[1] Quote retrieved from Bert Vaux's personal website (http://sitekreator.com/vaux/dialects.html) on May 10th, 2017.

linguistic background, with a focus on their regional dialect. At this point, some talkers will engage in some general meta-linguistic commentary (many will mention that they don't think they have an accent). Then talkers read a list of words designed to elicit dialect differences in pronunciation. The original list is made of up words which are shibboleths associated with at least one dialect of American English. For example, the word-list contains the items "water" (pronounced as [wʊɾɻ] in Pittsburgh) and "oil" (which can be pronounced as [al] in the South). Finally, listeners read and answer a list of questions designed to elicit lexical variation, such as "What is it called when you throw toilet paper on a house?" and "What is a bubbly carbonated drink called?".

This study focuses on only the word list portion of the accent tag, since different talkers have different answers in the lexical elicitation portion of the task. Over time, participants in the meme have changed and appended the word list, most notably with terms commonly used in online communities which have multiple possible pronunciations. Some examples of these terms are "GPOY" (gratuitous picture of yourself) or "gif" (graphics interchange format, a popular digital image format). Even with these variations, all videos discussed here used some subset of the word-list shown in Table 3.1.

Talkers A total of eighty talkers were sampled for this project. Videos for eight men and eight women from each dialect region were included. The dialect regions were California, Georgia, New England (Maine and New Hampshire), New Zealand and Scotland. These regions were chosen based on their high degree of geographic separation from each other, distinct local regional dialects and (relatively) comparable populations. Of these regions, California has by far the largest population, with approximately 38.8 million residents, and New England the smallest, with Maine and New Hampshire having a combined population of approximately 2.6 million (although the United States census bureau estimates the popu-

Aunt	Syrup	Pecan	Sandwich
Roof	Cool Whip	Marriage	Attitude
Route	Pajamas	Both	Officer
Wash	Caught	Again	Avocado
Oil	Catch	Probably	Saw
Theater	Naturally	Spitting	Bandanna
Iron	Car	Image	Oregon
Salmon	Aluminum	Alabama	Twenty
Caramel	Envelope	Guarantee	Halloween
Fire	Arizona	Lawyer	Quarter
Water	Waffle	Coupon	Muslim
Sure	Auto	Mayonnaise	Florida
Data	Tomato	Ask	Wagon
Ruin	Figure	Potato	GPOY
Crayon	Eleven	Three	Gif
New Orleans	Atlantic		

Table 3.1: Superset of words included in accent tag videos

lation of New England as a region at over 14 million as of 2010 (Bogue, Anderton, & Barrett, 2010)).

Sampling was done by searching YouTube using the exact term "accent challenge" or "accent tag" and the name of the geographical region. Only videos which had automatic captions were included in this study. For each talker, the word error rate (WER) was calculated separately by dividing the number of errors by the total number of words. So a WER of 1 would indicate that all words were recognized incorrectly, while a WER of 0 would indicate perfect performance.

3.2.0.2 Results

The effect of dialect and gender on WER was evaluated using linear mixed-effects regression. Both talker and year were included as random effects (both slope and intercept). Talker was included to control for both individual variability in speech clarity and also recording quality,

since only one recording per talker was used. Year was included to control for improvements in ASR over time; since automatic captions are generated just after the video is uploaded to YouTube, and the recordings used were uploaded over five year time span it was important to account for overall improvements in speech recognition.

A model which included both gender and dialect as fixed effects performed better than a similar model without gender (χ^2(5, N= 80) = 31, p $<$ 0.01), without dialect (χ^2(5, N= 80) = 14, p $<$ 0.01) or without either (χ^2(5, N= 80) = 31, p $<$ 0.01). In terms of dialect, talkers from Scotland had reliably worse performance than talkers from the United States or New Zealand, as can be seen in Figure 3.1. The lower level of accuracy for Scottish English can't be explained by, for example, a small number of talkers of that variety. The population of New Zealand, the dialect which had the second-lowest WER, is roughly 80% that of Scotland. Nor is it factor of wealth. Scotland and New Zealand have a GDP per capita that falls within one hundred US dollars of each other. This is relevant because Google collects voice data from users of their services in order to train and refine models, and there may thus be a higher error rate for talkers from poorer countries, which might have a slower rate of adoption of new technologies.

There was also a significant effect of gender: women had higher error rates than men (t(78) = -3.5, p $<$ 0.01). This is shown in Figure 3.2. While there is quite a bit of intertalker variation, overall women have a much higher word error rate than men. This is somewhat surprising given relatively recent studies which found the opposite result (Goldwater et al., 2010; Sawalha & Abu Shariah, 2013).

Given the nature of this project, there is limited access to other demographic information about the talkers which might be important, such as age, level of education, socioeconomic status or ethnicity. The last is of particular concern given recent findings that automatic natural language processing tools, such as language classifiers and parsers, struggle with

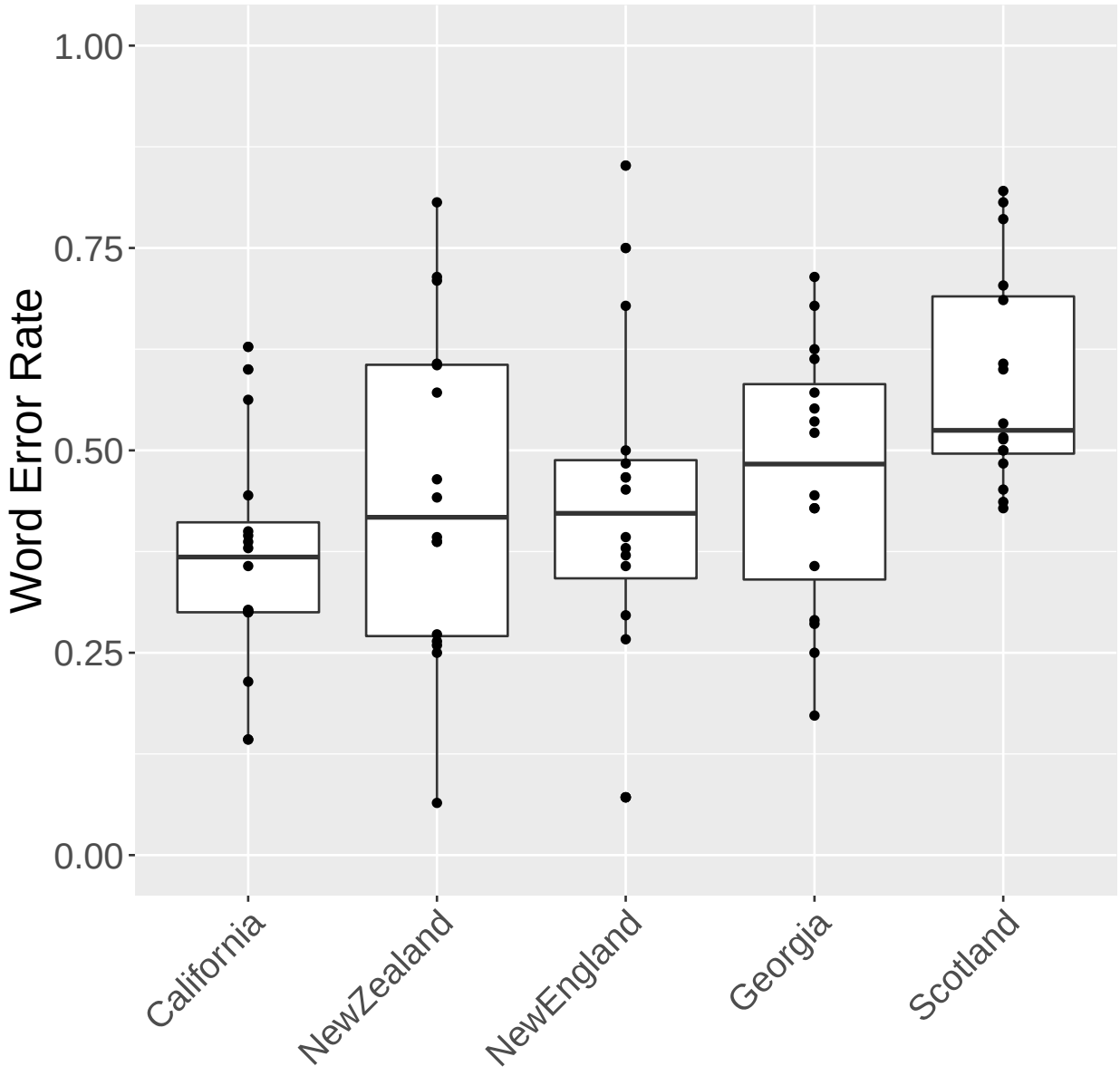

Figure 3.1: YouTube automatic caption word error rate by talker's dialect region or country.

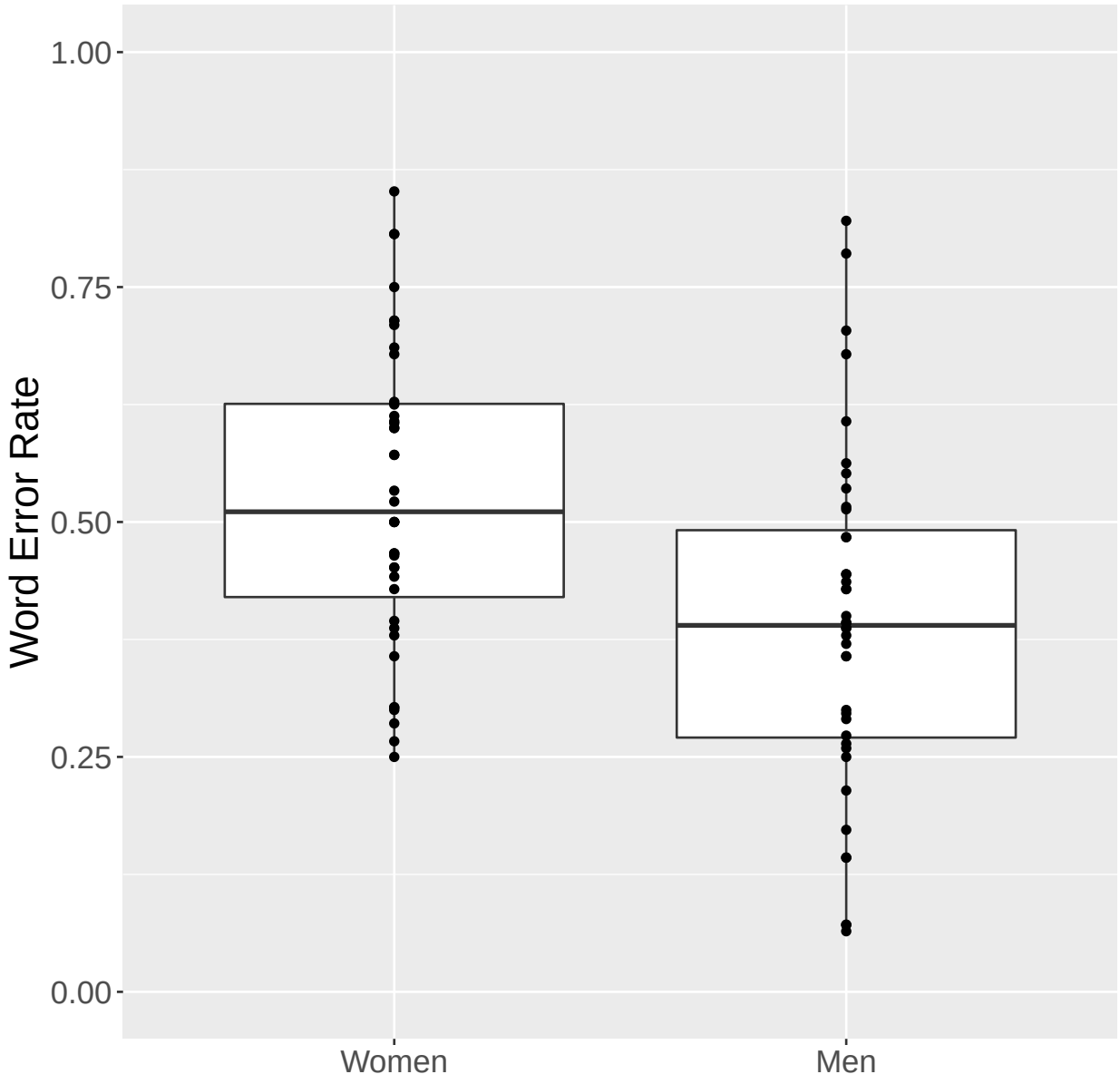

Figure 3.2: YouTube automatic caption word error rate by talker's gender.

African American English (Blodgett et al., 2016).

3.2.0.3 Effects of pitch on Youtube automatic captions

One potential explanation for the different error rates found for male and female talkers is that the talker's pitch is affecting the speech recognition algorithm. Pitch differences are one of the most reliable and well-studied perceptual markers of gender in speech (Wu & Childers, 1991; Gelfer & Mikos, 2005) and speech with a high fundamental frequency (typical of women's speech) has also been found to be more difficult for automatic speech recognizers (Hirschberg, Litman, & Swerts, 2004; Goldwater et al., 2010). A small experiment was carried out to determine whether pitch differences were indeed underlying the differing word error rates for male and female talkers.

First, a female talker of standardized American English was recorded clearly reading the word list shown in Table 3.1. In order to better approximate the environment of the recordings in the accent tag videos, the recording was made using a consumer-grade headset microphone in a quiet environment, rather than using a professional microphone in a recording booth. The original recording had a mean f0 of 192 Hz and a median f0 of 183 Hz, which is slightly lower than average for a female talker of American English (Pépiot, 2014). The pitch of the original recording was artificially scaled both up and down 60 Hz in 20 Hz intervals using Praat (Boersma et al., 2002). This resulted in a total of seven recordings: the original, three progressively lower pitched and three progressively higher pitched. These resulting sound-files were then uploaded to Youtube as a video and automatic captions were generated for the word lists. The video, and captions, can be viewed on Youtube at the following URL: https://www.youtube.com/watch?v=eUgrizlV-R4.

Overall, the automatic captions for the word list were very accurate; there were a total of 9 errors across all 434 tokens, for a WER of .002. Though it may be due to ceiling effects,

there was no significant effect of pitch on classification accuracy. The much higher accuracy of this set of captions may be due to improvement in the algorithms underlying the automatic captions or the nature of the speech in the recording, which was clear, careful and slow. More investigation with a larger sample of voices is necessary to determine if pitch differences, or perhaps another factor such as intensity, are what is underlying the differences in WER for male and female talkers.

3.2.0.4 Discussion

The results presented above suggest two things:

- there are differences in WER between dialect areas and genders
- manipulating one talker's pitch was not sufficient to affect WER for that talker

While the latter needs additional data to form a robust generalization, the large sample and effect size for the former is deeply disturbing. It begs the question: why do these differences exist? From a linguistic standpoint there is nothing inherently more or less recognizable about speech sounds in these different dialect areas, and the fact that earlier research has found lower WER for female talkers shows that creating such an ASR system is possible. It seems more likely, especially given that there is also a difference between dialects, that these differences are due to something besides the inherent qualities of the signal.

However, there are several shortcomings with the project presented here. First, it looked at accuracy for words spoken in isolation. Since it is standard for ASR systems to consider the linguistic context in which a word is spoken, only considering words spoken in isolation places the system at a disadvantage. Second, since the captions were generated over the span of several years, it does not provide a snapshot of a system at one point in time and may not be a good description of the system. Third, it only considered accuracy for a single system.

And, fourth, it does not include an acoustic analysis of the speech varieties being described in order to determine that they do represent robust regional variation. The second section of this chapter expands on the first by reevaluating YouTube's automatic captions using connected speech and includes an evaluation of a second system, Microsoft's Bing Speech API. Both systems are evaluated using the same recordings of connected speech that were transcribed by the systems over the period of a few minutes. It also includes an acoustic analysis of regional differences in the speech data used to evaluate the automatic speech recognition systems.

3.2.1 YouTube's automatic captions and Bing Speech

3.2.1.1 Four varieties of American English: Data

Four distinct varieties of American English were chosen for the second comparison: General American English, Northern Cities, Southern and California English. These are all major, well-documented varieties of American English which are acoustically distinct. For Californian English, only speakers from California were chosen. For the Northern Cities and Southern varieties, speakers from Michigan and Alabama, respectively, were used. An acoustic analysis was carried out on each group of speakers to ensure that they were participating in these vowel shifts. Some of differences between these varieties can be seen in Figure 3.3, and are discussed at greater length in the sections below.

Acoustic data for these varieties was taken from the International Dialects of English Archive (IDEA) (Meier & Muller, 1998), which was created by Paul Meier as a resource for actors. The archive contains over 1,300 recordings from more than 120 countries, representing a wide variety of talker backgrounds. Each recording includes a reading passage followed by a short sample of unscripted speech and is generally no more than four minutes long. Many of the recordings were made by Meier, and additional contributions are continuously

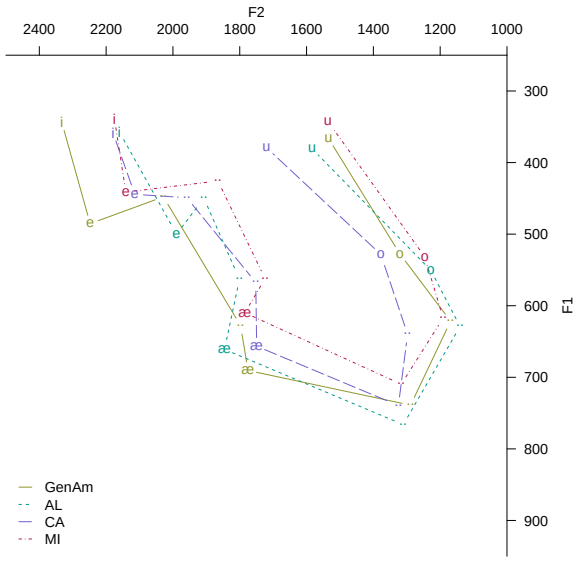

Figure 3.3: Figure showing the vowel space of acoustic data from General American, Alabama, California and Michigan talkers in an F1 - F2 space.

added by volunteer editors. The environments in which the recordings are made vary quite a bit, from field recordings made in talker's kitchens to recordings made in sound-attenuated booths. As a result, many of the recordings are of lower quality than those typically used in phonetic research. While this is less than ideal for a fine-grained acoustic analysis, it is actually a benefit for this study since it is better representation of the type of speech data a commercial ASR system might be expected to deal with. All speech data is taken from talkers reading the passage "Comma Gets a Cure" (Honorof, McCullough, & Somerville, 2000), which was designed to include Wells Standard Lexical Set (Wells, 1982), and is read by almost all talkers in the archive. (Very early recordings, where talkers read "The Rainbow Passage" instead are also of lower recording quality and are excluded from this study.) While this does mean that the text evaluated here is read speech, which does have different acoustic qualities than spontaneous speech (Warren & Hay, 2012), it also means that all talkers are

producing the same words. The latter is particularly important because it means that no talker will have an artificially high error rate due to their use of rarer words.

The acoustic dialectal differences observed in the data and discussed below shows that these are four distinct varieties of American English, and that they include phonetic features which are associated with these varieties in the sociophonetic literature. The strong differences between these varieties should provide a robust metric of the ability of automatic speech recognition (ASR) systems to handle dialectal variation.

A total of 39 talkers were included in this analysis: 11 from Alabama, 8 from California, 8 from Michigan and 12 General American talkers. There were slightly more male talkers than female talkers (22 men, 17 women). For 13 talkers, including all the General American talkers, their race was unreported. Among the remaining talkers, 13 were white, 8 African American, 4 of mixed race and 1 Native-American. A full table of talkers can be seen in Table 8.2 on page 151.

It is important to note that, unlike in the first YouTube evaluation, there is an imbalance of speakers across the demographic groups considered here. This is summarized in Table 3.2 on the following page. In particular, there are very different numbers of African American speakers across each state. The majority of African American speakers in this study are from Alabama, with an additional two from Michigan. Due to the small sample size, African-American talkers from both states are grouped together, which does not account for regional variation in African American English (Hinton & Pollock, 2000).

Mainstream US English, General American English Mainstream US English (MUSE), also called Standard[ized] American English and General American English, is the prestige variety of English in the United States. Unlike in Britain, where the prestige variety of Received Pronunciation has widely agreed-upon "correct" pronunciations, MUSE a collection

	African-American	Caucasian	Mixed	Native-American	Unknown
Alabama	6	7	0	0	0
California	0	5	3	1	1
Michigan	2	12	1	0	0
General American	0	0	0	0	12

Table 3.2: Distribution of speakers by ethnicity and state. Note that most African-American talkers were from Alabama, and that the ethnicity of the General American talkers was not reported and is thus unknown (although it is likely that the majority are Caucasian).

of American English varieties characterized by their lack of stigmatized linguistic features (Wolfram & Schilling, 2015). The most stigmatized features in American English are lexical (e.g. "ain't", "skeeter") and morphological (e.g. double negation, double modals). However, some phonological features are also stigmatized. In particular, talkers who are participating in major, on-going vowel shifts, especially in the South, may be regarded as "less correct" by other talkers (Preston, 2002). From the perspective of a linguist, however, no one dialect or language variety is "more correct" than any other.

The MUSE speech samples used in this project were produced by voice professionals (actors, voice coaches and speech language pathologists) who consciously avoided using stigmatized features. In particular, none of the talkers included in this sample are participating in any of the three on-going vowel shifts in the US: the Northern Cities, Southern and California vowel shifts. This can be seen in Figure 3.4. Note that, compared to other talkers, the MUSE women in participial showed a larger vowel space. This is probably due to these talker's high degree of hyper-articulation (Johnson, Flemming, & Wright, 1993), which is not surprising given that they have all received explicit training in clear speech. As can be seen in Figure 3.4, male MUSE talkers do show a merger between \ɔ\ and \ɑ\. This does not indicate that they are not MUSE talkers, however, as this merger is not stigmatized.

Given that these talkers are both intentionally speaking clearly and also not users of

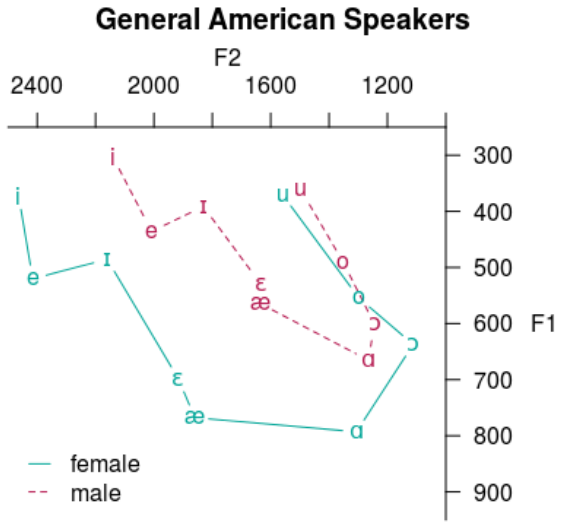

Figure 3.4: Figure showing the vowel space of acoustic data from General American talkers in an F1 - F2 space, separated by gender. Note the larger vowel space relative to other varieties.

stigmatized phonetic features (which may be underrepresented in the training data), it seems likely that ASR systems will transcribe their speech more accurately than that of talker using other dialects.

Northern Cities Vowel Shift The Northern Cities vowel shift is an on-going vowel shift in the North East and Northern Midwest, especially in major cities in Michigan and Illinois, including Chicago (Labov, Yaeger, & Steiner, 1972). It is characterized by lowering of \ɪ\ to \ɛ\, backing of \ɛ\ to \ʌ\, backing of \ʌ\ to \ɔ\, lowering of \ɔ\ to \ɑ\, fronting of \ɑ\ to \æ\ and raising of \æ\, sometimes as high as \i\ (Gordon, 2001). The Michigan talkers included in this study are participating in parts of this on-going chain shift, as can be seen in Figure 3.5 on the next page. In particular, \æ\ is raised past the General American \ɛ\ and \ɛ\ is slightly backed.

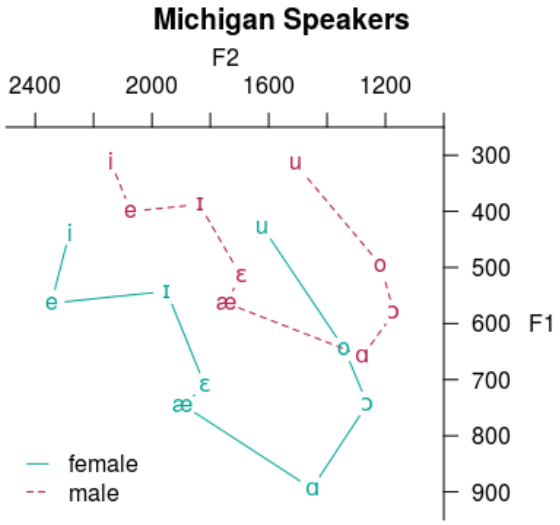

Figure 3.5: Figure showing the vowel space of Michigan talkers in an F1 - F2 space, separated by gender. Note the raising of \æ\ and backing of \ɛ\.

Southern Vowel Shift The talkers from Alabama have phonetic features of Southern English, including both participating in the Southern Shift and maintaining a strong distinction between \ɔ\ and \ɑ\. The components of the Southern Shift most apparent in these talkers are the raising and fronting of \ɪ\ and \ɛ\ and the lowering and backing of \e\ (but not \i\, as has been observed in other talkers (Wolfram & Schilling, 2015)). In addition, both \u\ and \o\ are fronted. Though the formant measurements shown in Figure 3.6 were taken at the midpoint of vowels and thus do not show this, the Alabama talkers also had a high degree of \ai\ monophthongization, a feature found in both white and Black Southern American English (Fridland, 2003).

California Vowel Shift The Californian talkers in this sample are participating in the California Vowel Shift (Eckert, 2008), including fronting of \u\ and \o\, backing of \ɑ\ towards \ɔ\ and lower and retraction of \ɪ\ and \ɛ\.

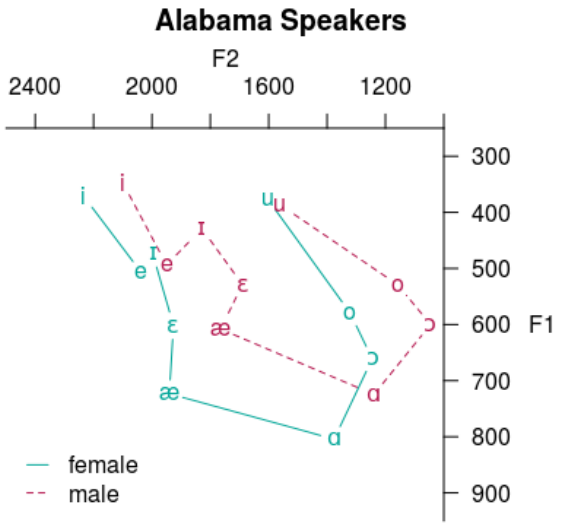

Figure 3.6: Figure showing the vowel space of talkers from Alabama in an F1 - F2 space. Note the \ɪ\ is raised while \e\ is lowered and \ɛ\ is raised.

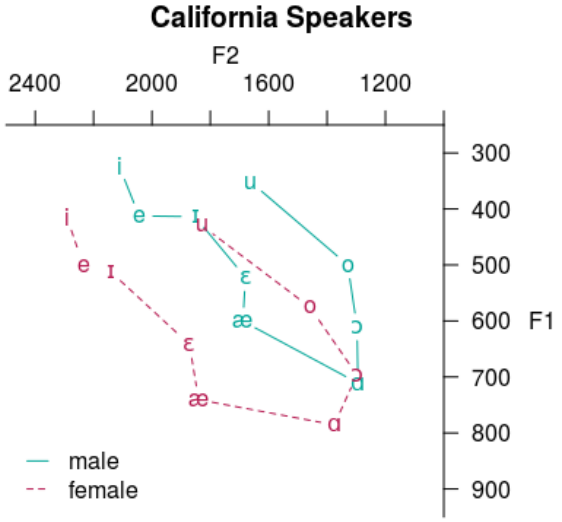

Figure 3.7: Figure showing the vowel space of acoustic data from California in an F1 - F2 space. Note that both \u\ and \o\ are very fronted.

3.2.1.2 System evaluation

Bing Speech (Project Oxford) All of the speech files discussed above were transcribed using the Bing Speech API [2], a commercial speech recognition product offered through Microsoft Cognitive Services, previously Project Oxford. The API was accessed through a custom Android Application built for this analysis and developed using the Bing Speech Android SDK [3]. The app sent each file to be transcribed individually. In order to improve efficiency, the .wav files were down-sampled from 22050 Hz to 11025 Hz. The returned transcriptions were stored in a separate .txt file for each audio file. Not all files were transcribed, and of those that were, many had truncated transcriptions. Of the 48 files sent to the API, only 36 were returned with transcriptions, despite repeated efforts. It is unclear why this happened. Many possible hypothesis were checked–including the possibility that files were being transcribed up to a certain time point, that only a set amount of audio data was being transcribed, that transcriptions were returned once a complete syntactic structure was formed–and none could account for which files' transcriptions were truncated and where. It is possible that transcriptions were returned only up until the point where a running measure of confidence in the transcription dropped below a certain threshold, but this could not be verified. It is also worth noting that the API, at the time that this study was conducted, was newly-launched and that the return of incomplete transcriptions might be the result of a bug.

YouTube automatic captions Audio files were also transcribed using YouTube's automatic captions (Harrenstien, 2009). These were generated by creating MP4 videos with the audio files as a soundtrack. The language was manually set to "English (United States)".

[2] https://www.microsoft.com/cognitive-services/en-us/speech-api

[3] https://github.com/Microsoft/Cognitive-Speech-STT-Android

Once generated, the automatic captions were then downloaded as .srt files and converted to plain text prior to analysis. The conversion to plain text did result in the removal of information about ASR recognition confidence, which is color-coded by word in automatic captions.

Word error rate Because many of the transcriptions returned by Bing were partial, calculating Word Error Rate (WER) as the by-word edit distance from the correct transcription would have led to an artificially high WER. (But see (Y.-Y. Wang, Acero, & Chelba, 2003) for a broader discussion of the shortcomings of WER). To correct for this, the WER was calculated as the number of non-deletion errors divided by the total number of words in the automatic transcription. So if the correct transcription was “The lamb is cute" and the returned transcription was “The lamb shoots", the WER would be 0.33 (one substitution over three words) rather than 0.5 (one substitution and one deletion over four words).

Overall, the WER were quite high, especially given very high accuracy (under 0.07) recently reported by a team at Microsoft (Xiong et al., 2016). The mean WER for the Bing transcriptions was 0.45 ($\sigma = 0.18$). The highest error rates were for those files where only a few words--in one case as few as two--were transcribed and those incorrectly. The YouTube WER was both lower and less variable ($\mu = 0.31, \sigma = 0.07$).

Difference in accuracy by dialect WER did vary across dialects, as can be seen in Figure 3.8. For both systems, the lowest average WER was for General American talkers and the highest for talkers from California. While the former was expected, the latter is surprising given the finding above that YouTube’s automatic captions had the highest accuracy for Californian talkers. Differences in WER by dialect were not robust enough to be significant for Bing (under a one-way ANOVA) ($F[3,32]=1.6$, $p = 0.21$), but they were for YouTube’s automatic captions ($F[3,35] = 3.45$, $p < 0.05$). The differences between these two systems is

not surprising given the far lower variance for the YouTube WER.

Difference in accuracy by talker gender The previous findings of an effect of talker gender on WER was not replicated here for either system. Neither Bing ($F[1,34]=1.13$, $p = 0.29$), nor YouTube's automatic captions ($F[1,37] = 1.56$, $p = 0.22$) had a significant difference in accuracy by gender.

This was true for both the General American talkers and talkers from other dialects. Even when talkers were split into General American and non-General American groups, there was no group which showed a significant effect for talker gender for either Bing or YouTube's automatic captions.

Difference in accuracy by talker race As can been seen in Figure 3.11, for both systems, error rates were lowest for white talkers as a group, and higher for African American and mixed race talkers. As with dialect, differences in WER between races were not significant for Bing ($F[4,31]=1.21$, $p = 0.36$), but were significant for YouTube's automatic captions ($F[4,34]=2.86$, $p<0.05$).

3.2.1.3 Discussion

In the second half of this chapter, I evaluated two automatic speech recognition systems, Microsoft's Bing Speech and Google's YouTube automatic captions, on a sociolinguistically-stratified sample of talkers from different dialect backgrounds, genders and races. While both systems made errors, the rate at which errors were made varied based on talker's social identities, and in particular their dialect background and race. These results were only statistically reliable for YouTube's automatic captions, although given the high variability in Bing's WER, the sample size was too small to achieve high power. Given four dialect regions, an F of .35 (as observed for the YouTube captions) and a significance level of 0.05,

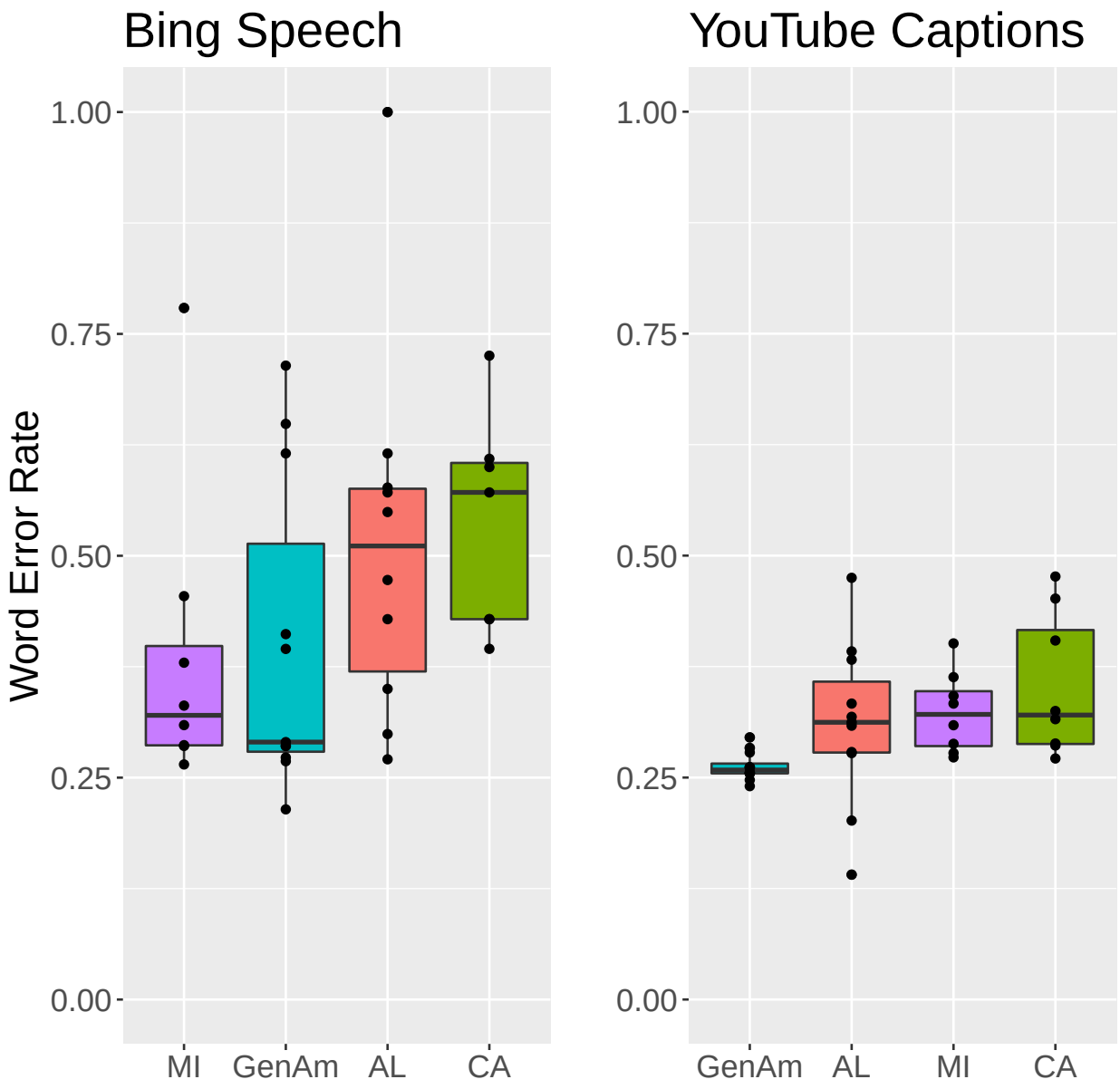

Figure 3.8: Word error rate by region. Points represent individual talkers.

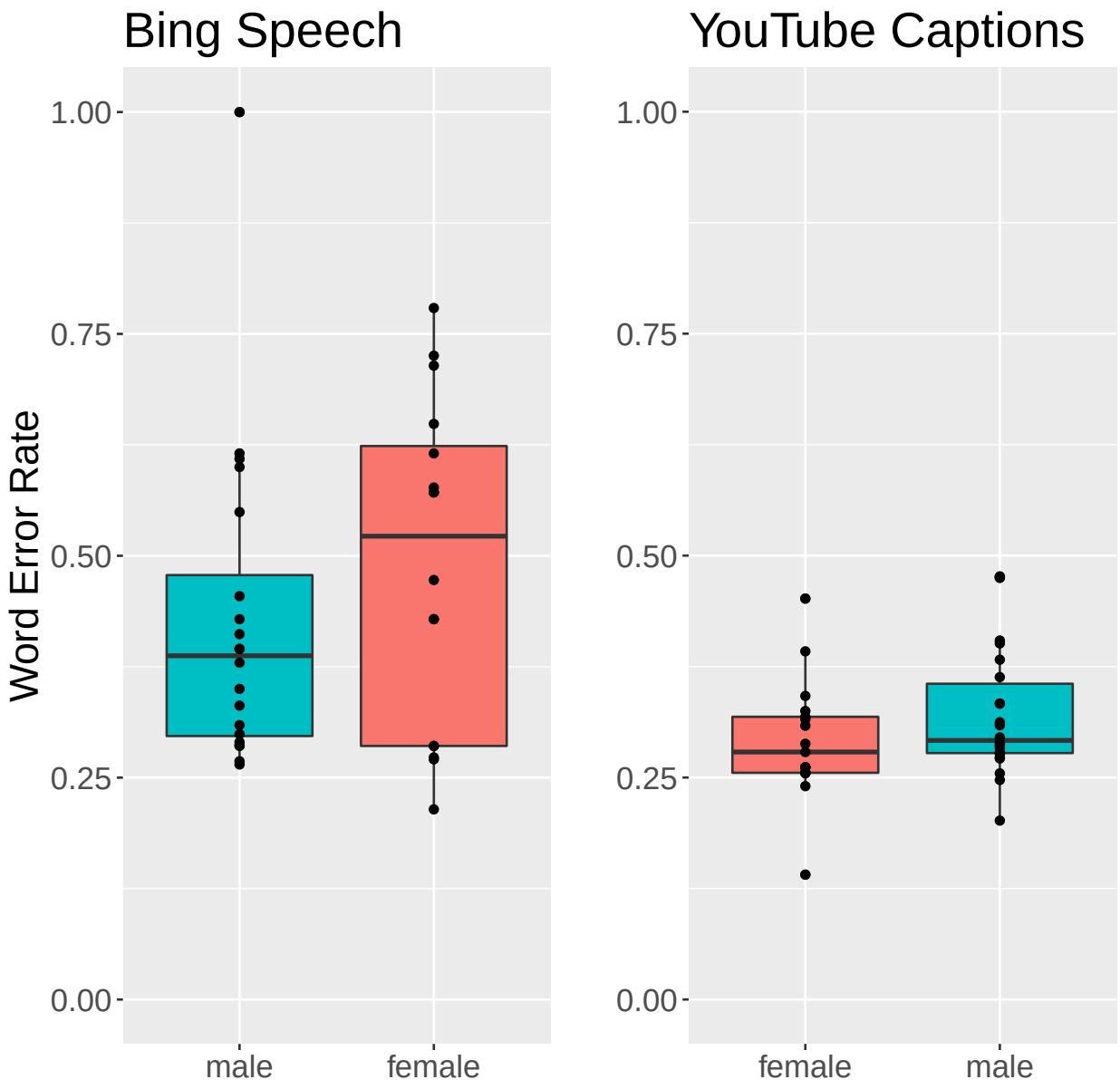

Figure 3.9: Plot showing word error rate by gender. Points represent individual talkers.

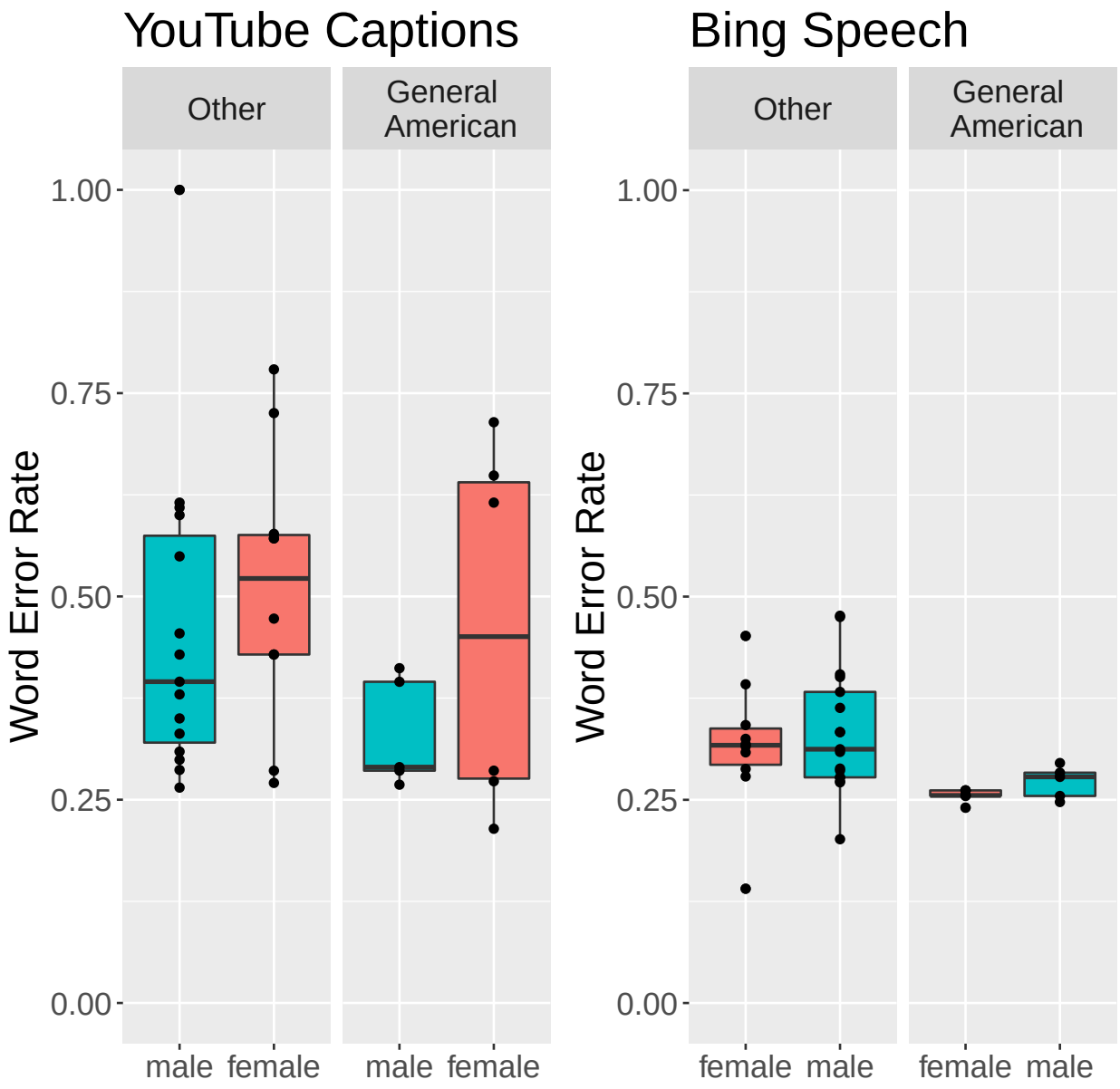

Figure 3.10: Word error rate by gender and system, with General American speakers separated from other speakers.

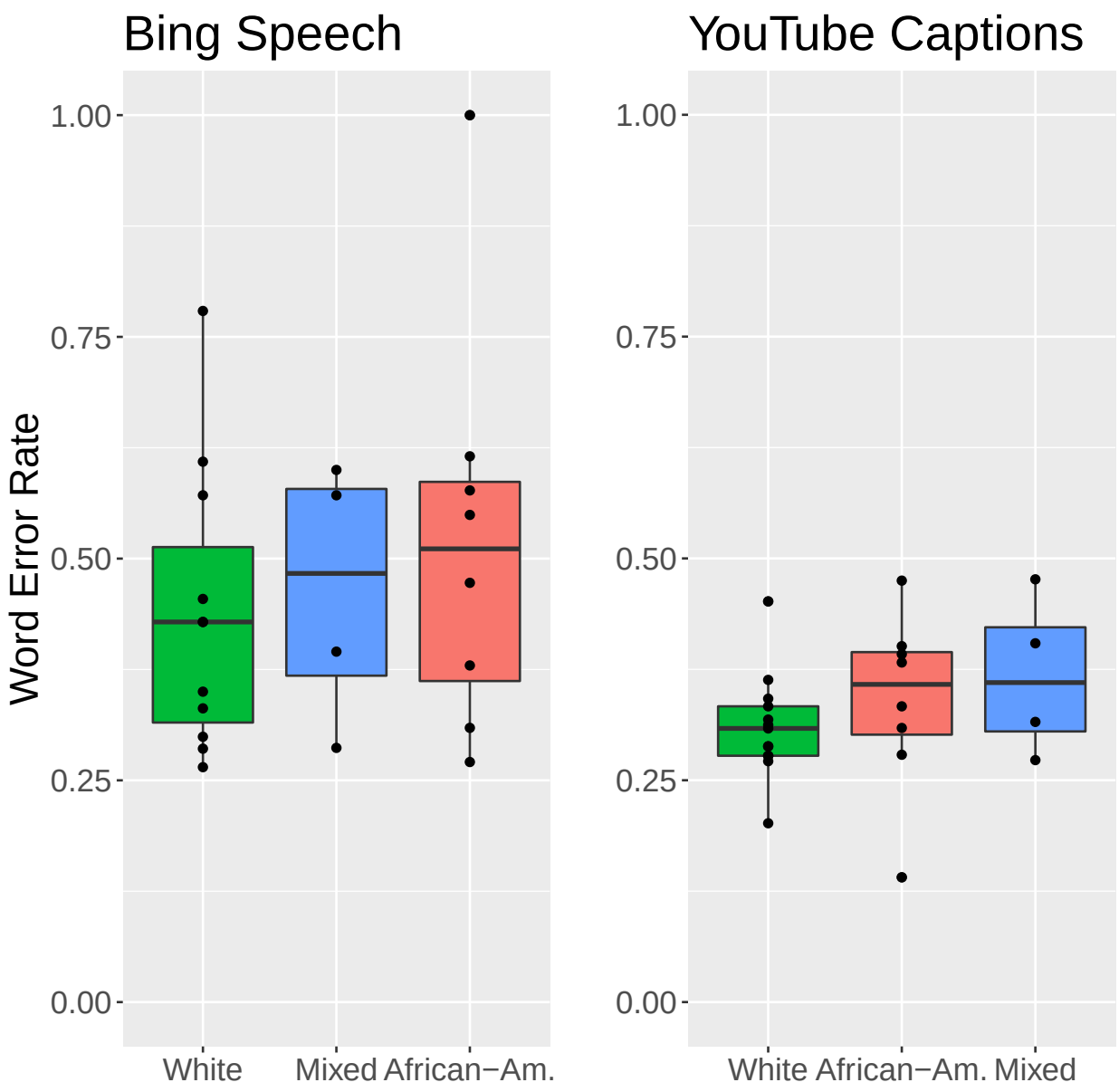

Figure 3.11: Plot showing word error rate by region and talker race (excluding talkers of unknown race and the one Native American talker). Points are individuals, and bars are ordered left to right from the lowest to highest average WER.

at least 24 talkers per dialect should be sampled to obtain a power of 0.8 (Champely, 2016). It was not possible to increase the sample size, however, given that all talkers in the archive who read “Comma Gets a Cure" (Honorof et al., 2000) from each state were included.

However, even with the small sample size, some robust effects of talker ethnicity and dialect were observed, and the direction of the effects was the same across systems. Among the dialects, both systems had the lowest average WER for General American talkers, and among ethnicities, both systems had the lowest WER for white talkers. The former, of course, is possibly confounded by the fact that the General American talkers in this study were all voice professionals. As a result, they produced very clear, hyper-articulated speech. It is possible that the differences between dialects arose from this rather than a bias towards a particular dialect. This would not, however, account for biases towards white talkers, as the ethnicity of the General American talkers was unknown and they were thus excluded from racial/ethnic analysis.

Why was accuracy so much higher for General American talkers than others? One potential reason, as discussed above, is the fact that the General American talkers in this sample were voice professionals and the quality of their recordings was generally higher than it was for other speakers. Another potential reason is that General American speakers may be over-represented in the data used to train these systems. General American is associated with white, educated, upper middle class speakers, which is also the population most likely to have earlier access to new technologies incorporating ASR. This is an important consideration, since users’ voice data is fed back into systems as training data (Schalkwyk et al., 2010). As a result, early over-representation of talkers from these demographics in the userbase may result in over-fitting to the language varieties used by the these talkers, at the expense of accuracy for speakers from different speech communities.

3.2.2 What causes differences in accuracy between different social groups?

Both evaluations found that listener's social identities was tied to differences in accuracy in word recognition, although which aspects of talkers' identity mattered varied between evaluations. So what might lead to these differences?

One candidate for the cause of these differences is imbalances in the training dataset. Any bias in the training data will be embedded in a system trained on it (Torralba & Efros, 2011; Bock & Shamir, 2015). While these ASR systems are both propriety and it is thus impossible to validate this supposition, there is room for improvement in the stratification of many speech corpora. Librevox, for example, is a popular open-source speech data set that "suffers from major gender and per talker duration imbalances" (Panayotov, Chen, Povey, & Khudanpur, 2015). TIMIT, the most-distributed corpus available through the Linguistic Data Consortium, is balanced for talker dialect but approximately 69% of the speech in it comes from male talkers (Garofolo et al., 1993). Many other popular speech corpora such as the numbers corpus (Cole, Noel, Lander, & Durham, 1995) or the AMI meeting corpus (McCowan et al., 2005) don't include information on talker gender or dialect background in the papers introducing the resources. Taken together, these observations suggest that socially stratified sampling has historically not been a priority during corpus construction for computational applications.

One solution, of course, is to focus on collecting unbiased socially stratified samples, or at least documenting the ways in which samples are unbalanced, for future speech corpora. This is already being addressed in the data collection of some new corpora such as the Automatic Tagging and Recognition of Stance (ATAROS) corpus (Freeman et al., 2014).

This does not help to address existing imbalances in the data, however. One way of doing this is to include information about talker identity as a feature during speech recognition in order to help capture the stable variation between groups of talkers. This may be done,

by, for example, including the geographic location of the talker (Ye et al., 2016), gender-dependent speech recognition models (Abdulla, Kasabov, & Zealand, 2001) or models trained on specific ethnolects (Lehr, Gorman, & Shafran, 2014). Another solution to this problem is an application of the research presented in this dissertation: automated accent adaptation. While adapting to novel language varieties remains a problem for large-scale ASR, human listeners are able to quickly adapt to and correctly perceive language in the presence of stable variation. Modeling this process can provide specific recommendations for improving automated accent adaptation, and this is the focus of the rest of this dissertation.

3.3 Is dialect adaptation for ASR necessary?

Yes, there is still a need for dialect adaptation for automatic speech recognition. As it stands, a speaker's linguistic variety, which is a reflection of their social identity, directly affects how accurate state-of-the-art commercial automatic speech recognition systems work for them. The good news is that dialect bias in ASR is a fixable problem: all dialects are internally consistent, systemic and thus learnable. The challenge is to make that learning fast, accurate and robust and then deploying it in a way that does not reduce accuracy for any group of speakers.

Chapter 4

CAN SOCIAL INFORMATION OUTWEIGH RECENT PERCEPTUAL LEARNING?

4.1 Background

Perceptual learning is the process by which repeated exposure to stimuli results in a long-term change in the perceptual system. In terms of speech perception, this usually takes the form of either learning to re-categorize a speech sound–perhaps learning to perceive /l/ and /ɹ/ as different sounds, or merging two previously distinct vowels–or in changing the attentional weight to certain parts of the acoustic signal–as in cue trading (Samuel & Kraljic, 2009). One specific type of perceptual learning occurs when a talker is exposed to a dialect of their native language. This is of interest for two reasons. The first is that it may result in an individual who produces one set of dialect features, but experiences no difficulty in perceiving others–the "fluent listener" (Sumner & Samuel, 2009). Such individuals are of interest to researchers working on variation in perception. This is especially true given that, on a neurological level, perception of different dialects differs measurably from perception of foreign languages or distorted speech (Adank et al., 2015). In addition, how best to adapt to novel dialects with an existing system is a concern in automatic speech recognition (ASR) (William et al., 2013; Najafian, DeMarco, et al., 2014, for some recent work). Modeling how humans accomplish the same task offers interesting and potentially useful insights for ASR researchers.

Within the perceptual learning literature, one usually unstated and previously untested question emerges: Do talkers rely on social information during perceptual learning? We

know that even very subtle social cues can have a strong effect on speech perception in general (Niedzielski, 1999; Koops et al., 2008; Drager, 2010; McGowan, 2015). But how does information about the demographics of a talker influence a listener's perceptual learning and determine when they choose apply that learning? Other researchers have certainly argued that it should (Kleinschmidt & Jaeger, 2015; Dahan et al., 2008), but less experimental work has directly investigated this question.

There is evidence that listeners can generalize recent perceptual learning to novel talkers and that this is mediated by social information. For example, there is evidence that the dialect of the experimenter who meets participants prior to their participation in a perception experiment influences their performance in that experiment (Hay, Drager, & Warren, 2010). In addition, greater variability in the training data–especially hearing more talkers–gives listeners an advantage in perceptual learning (Logan et al., 1991; Bradlow & Bent, 2008). This can be very intuitively explained by including social categorization as an integral part of perceptual learning: if talkers hear more talkers from a given social category, they have a more robust perceptual picture of that category. There also seems to be some benefit to intentionally asking participants to focus on talker identity: when listeners were presented with talkers who they had previously been exposed to, explicit cues about the talker's identity boosted perception (Trude & Brown-Schmidt, 2012).

However, listeners do not always carry over their perceptual learning effects. One study found that if listeners were trained on one voice they then showed improved accuracy when hearing the same voice in noise–a perceptual learning effect sometimes referred to as the "familiar talker advantage" (Levi et al., 2011). However, talkers who were trained on a voice speaking German showed no advantage when listening to that same voice speak English. This suggests that listeners were not making use of the perceptual learning they had done on that voice, even if they had become familiar enough with it to almost always identify

the talker correctly. Of course, this might be explained as an effect that only applies across different languages. However, related work entirely on Dutch (Eisner & McQueen, 2005) found a similar effect. Listeners did show carry-over of perceptual learning when the same ambiguous fricative tokens used in the training phase were spliced into a novel talkers–even if the novel talker was male and the training talker was female. However, when an entirely novel voice with the same type of ambiguous tokens was used in testing, the perceptual learning effect did not carry over. Instead, listeners categorized tokens as if they were produced in their own dialect. Finally, earlier work that suggests that some perceptual learning effects are talker-specific (Nygaard et al., 1994).

On the face of it, these two types of studies seem to show contradictory effects. Listeners can generalize perceptual learning to new talkers, but sometimes they don't. This can't entirely be explained as talker-specific models: if perceptual learning were only talker-specific then the dialect of the researcher should have no effect, and once trained on a voice the advantage should always carry over. Sometimes listeners do generalize from one talker or set of talkers to another, and sometimes they do not. I believe that this can be explained by using a mechanism that is already well-studied in speech perception: the fact that listeners make social judgments about the talkers they are hearing. If they believe that the training and testing talkers are members of the same social group then they will extend perceptual learning. However, if they believe that the testing talker has a distinct social identity, they will not apply the new percepts they have developed.

4.2 Methods

The accent chosen for training in this experiment is New Zealand English (NZE). NZE has several advantages. The first is that it is a "naturally occurring" dialect of English. While some studies have successfully used "artificial" dialects, for example when looking at toddler's

perceptions (White & Aslin, 2011), using tokens taken from an existing accent ensures that the phonological system is one that learners are capable of acquiring. The second advantage is that, though NZE is a stable dialect of English, it hasn't achieved wide cultural saturation in America. Further, relatively few Americans have first-hand experience with New Zealand English. From 2010 to 2015, New Zealand had an average of 10,000 visitors from the United States annually (*International Visitor Arrivals to New Zealand: September 2015*, 2015). Compare that to the 576,600 American visitors to Australia from June 2014 to June 2015 (*International Toursim Snapshot*, 2015), and it is clear that far fewer Americans have had significant in-person exposure to NZE. As a result, it should be easier to find talkers of American English who are unfamiliar with the variety. Finally, and perhaps most critically to this investigation, NZE is undergoing a vowel shift. As a result, some tokens may be potentially ambiguous depending on a listener's judgment of the talker's social identity.

4.2.1 New Zealand vowels

The New Zealand vowels are undergoing a shift that makes them ideal for this study. The TRAP and DRESS vowels ([æ] and [ɛ] in American English) are both very raised, resulting in precepts (for non-fluent listeners) like "pit", or even "peat", for "pet" and "pen" for "pan". The KIT vowel [ɪ] is also very centralized, resulting in merger between "women" and "woman", with both produced closer to [wʊmʌn] (Hay, Maclagan, & Gordon, 2008). This chain shift of the short (or lax) front vowels is part of an ongoing shift in NZE relative to American English. It began with the raising of the TRAP vowel, which in turn pushed the DRESS vowel higher, which finally displaced the KIT vowel, which retracted and lowered to the center of the vowel space (C. I. Watson, Maclagan, & Harrington, 2000). Currently, the DRESS vowel is continuing its upwards trajectory and beginning to crowd the FLEECE vowel–in some cases even surpassing it in frontness (Maclagan & Hay, 2007).

The result is a set of lexical items which are potentially ambiguous to a trained listener of both American English and New Zealand English. For example, [bɪt] could either be the New Zealand English "bet" or the American English "bit". If the training mentioned above is carefully controlled perceptual learning, this allows us to pick apart how much of this perceptual learning is tied to beliefs about the talker's identity. If, even having just been trained to recognize New Zealand English, American English listeners revert to their native judgments when they believe they are listening to another American English talker, this is strong evidence that perceptual learning is dependent on social information.

4.2.2 Stimuli

Stimuli were tokens produced by two female native talkers of New Zealand English between the ages of 20 and 21 who were university students[1]. In the original tokens, vowels were produced in the H_D context in isolation (C. Watson, 2014). To create tokens , the flanking consonants were removed and only a duration-normalized 150 ms center slice of the vowel was used in the experiment. This was done in order to remove the duration contrast, which is strong between New Zealand vowels (Langstrof, 2009). The average formant values of each vowel, by talker, are shown in Figure 4.1. While resynthesized vowel tokens would have allowed for greater control of the acoustic characteristics of the tokens, resynthesis proved impractical due to the high degree of creak used by both talkers.

4.2.3 Experimental software

The experiment was administered on-line using the Psytoolkit experimental testing software version 2.0.4 (Stoet, 2010). The code used to run the experiment is available on the author's github page (link: github.com/rctatman/psyToolkit_perceptualAdaptation-).

---

[1]The author would like to extend her deep gratitude to Dr. Catherine Watson for generously providing speech data.

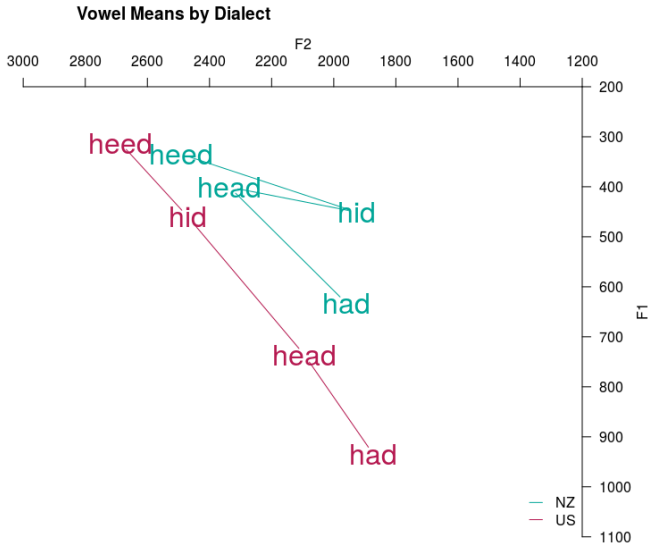

Figure 4.1: Vowel space of New Zealand English talker and American English talker. This figure suggests why NZE may be confusing for US English talkers; the NZE "head" tokens are very close to the position of "hid", and the "had" tokens are very close to the US "head" (although not precisely overlapping; these values are not normalized).

4.2.4 Paradigm

The experiment was divided into two stages: training and testing. Participants were trained on one talker and tested on the other, with each talker serving both roles. All participants were explicitly told that the first talker was from New Zealand, while half were told that the second talker was from the US and half were told that they were also from New Zealand. During each phase information was collected on participants' responses, the correct response and response times (although due to variation in computer hardware and Internet connections the latter should be treated as approximate). The overall structure of the experimental tasks can be seen in Figure 4.2.

Training During training, participants heard a vowel token and were asked to decide which word it had been taken from: "heed", "hid", "head" or "had". They were then given feedback on their choice (correct or incorrect) and the word that the token had actually been taken from. 100 tokens were presented, approximately thirty from each word. Note that, although "hid" was a choice, no vowel tokens were actually taken from the word "hid", due to its very short duration. Participants were passed on to the testing phase when they correctly answered eight in a row or after listening to one hundred tokens, whichever came first.

Testing Before the testing phase, participants were told whether the talker that they were going to be hearing was from New Zealand or the US, and shown the flag of the corresponding country [2]. They then listened to and categorized stimuli in the same way they had in the first training phase, but without receiving feedback.

[2]For New Zealand, the official flag as of November 2015 was used.

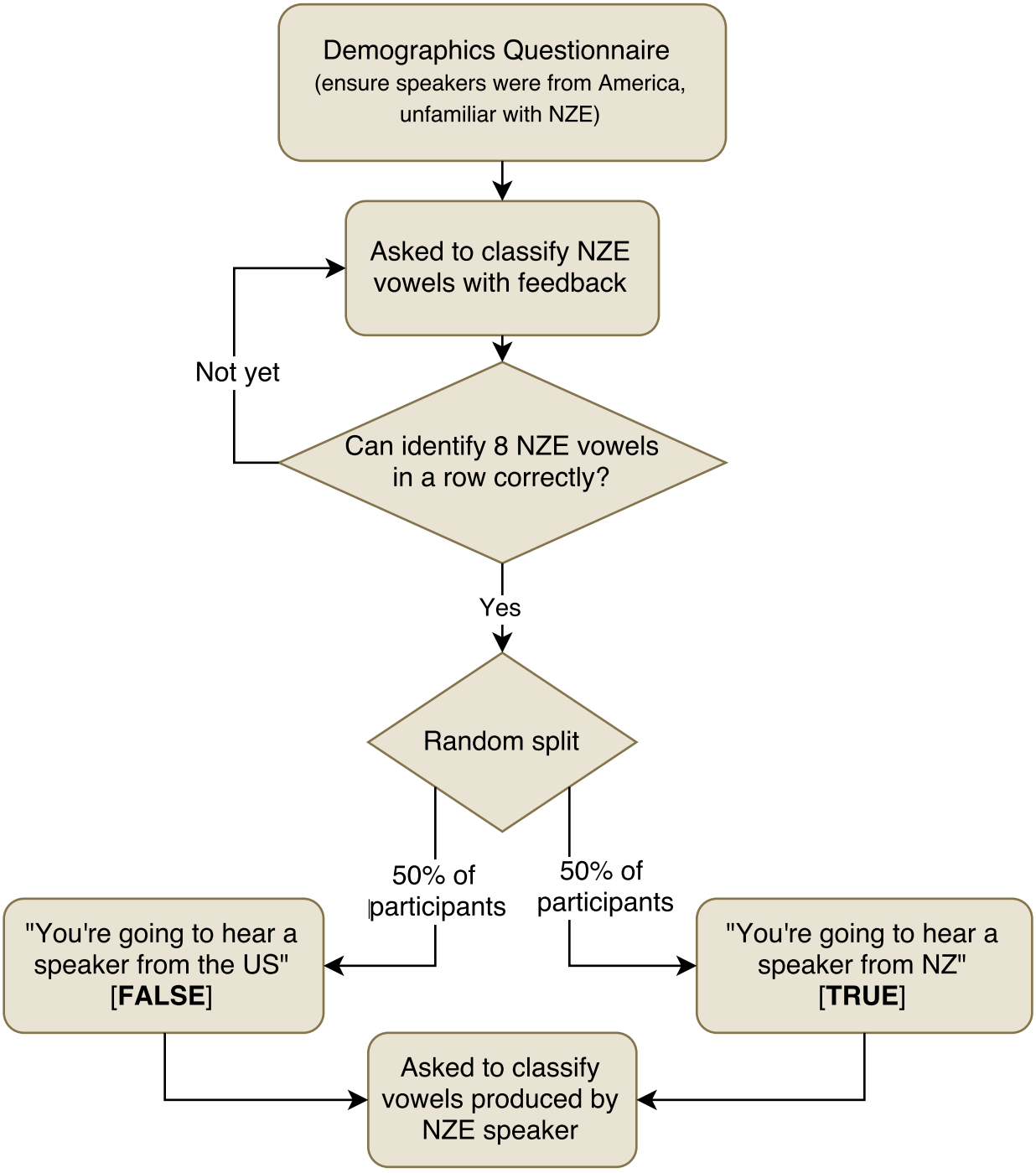

Figure 4.2: Flowchart depicting experimental design.

4.2.5 Participants

Fifteen native English talkers from North America were recruited online. (While twenty five participants began the survey, only fifteen completed the experiment.) Of these fifteen, ten were male, four were female and one was another gender. All were native English talkers from the United States who had never traveled to New Zealand and were either completely unfamiliar with or only somewhat familiar with New Zealand English. The age of participants ranged from 21 to 43, with a median of 25.

4.3 Results

The first question is whether the participants could successfully learn to identify the NZE vowels; if no perceptual learning took place, then it is a moot point whether it transferred to a second task. Participants could not move on from the first training block until they correctly classified eight tokens in a row. This took different participants between 91 and 9 trials, with a mean of 42 and a median of 40. No participants required all one hundred trials in order to learn to correctly classify the vowels.

Further evidence that participants did undergo perceptual learning comes from a confusability matrix, which is presented in Figure 4.3 on the following page. Note that while there is a fair amount of confusion, especially for "head" tokens, most tokens were classified correctly. This can be read off the matrix by the larger numeric values running along the diagonal. Data from the confusion matrix can be summarized statistically using Cohen's kappa (Cohen et al., 1960). A kappa value of 1, the maximum possible, would indicate no errors at all, while a value of 0, the minimum possible, would indicate that every single classification is an error. The overall kappa value for all training trials is 0.61, which is commonly characterized as either "substantial" (Landis & Koch, 1977) or "fair to good" agreement (Fleiss, Levin, & Paik, 2013).

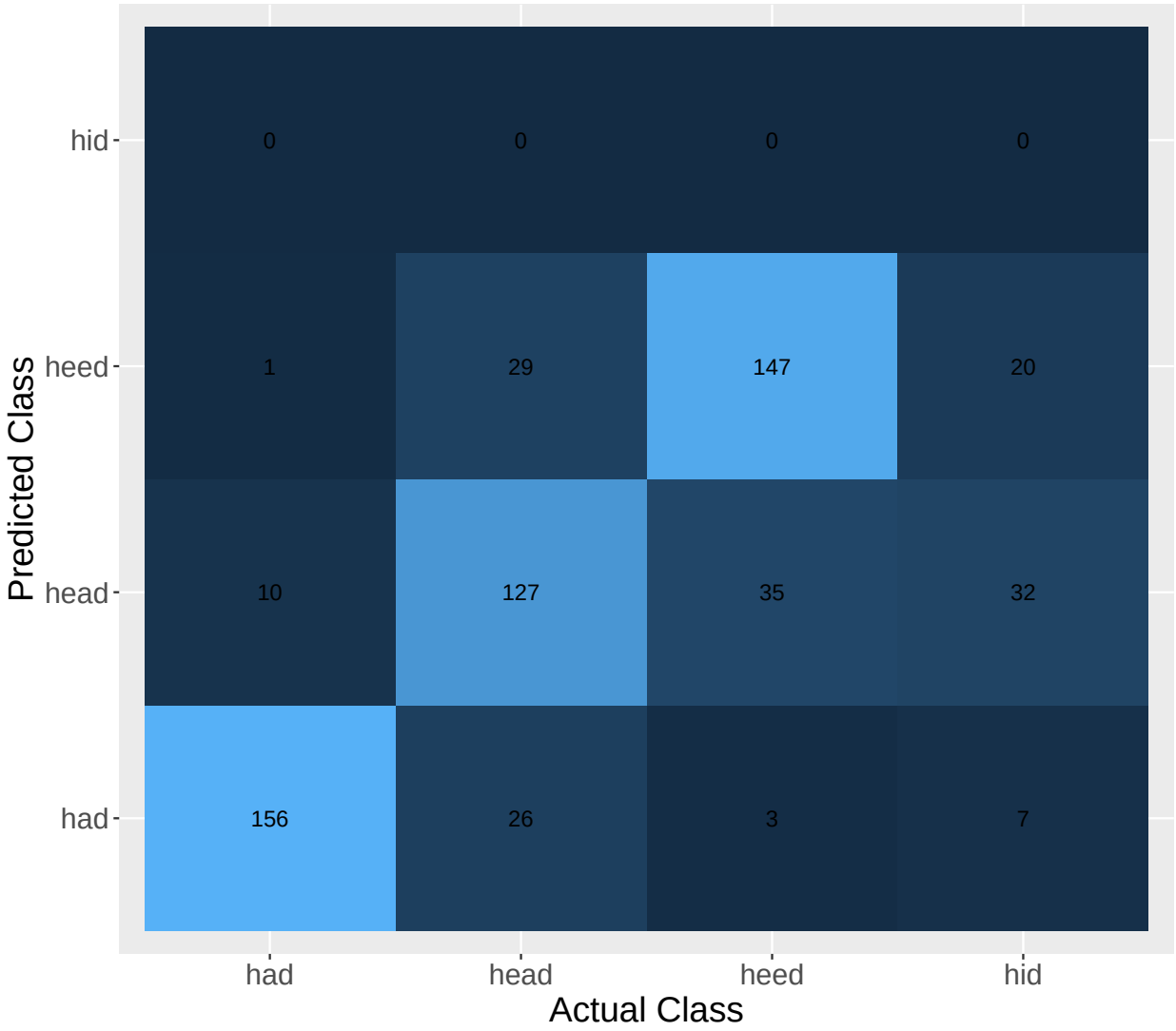

Figure 4.3: Confusability matrix based on all training trials. The column is the actual identity of the token, while the row indicate what participants responded. So the first column can be read, "Of all the 'had' tokens, 156 were correctly classified as 'had', 26 were incorrectly classified as 'head', 3 were incorrectly classified as 'heed' and seven incorrectly classified as 'hid'." The column for "hid" is empty because no actual tokens of "hid" were played for participants, though it was given to them as a possible classification.

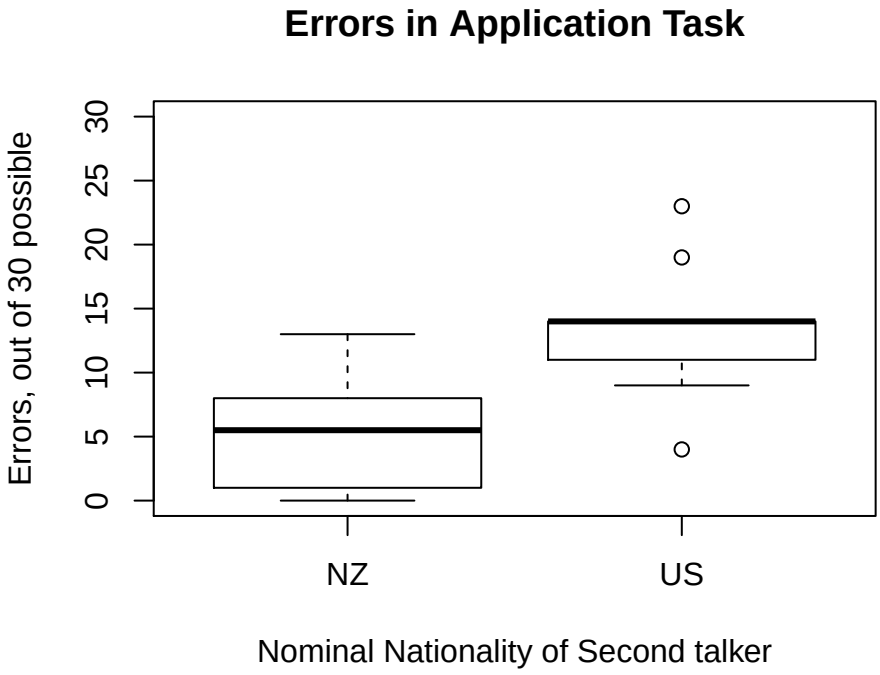

Figure 4.4: Classification errors during the second task, separated by the 'nationality' of the second talker. Note that listeners who believed they were listening to a talker from the US had many more classification errors than those who believed they were listening to one from NZ.

The second question that arises is whether there is a difference in response between the groups which were told (correctly) that the second talker was also from New Zealand and those who were told (incorrectly) that the second talker was from the United States. As we can see by looking at the total number of errors during the application task in Figure 4.4, it does appear that there's an overall difference, with listeners who were aware that they were listening to New Zealand English making many fewer errors. This difference in number of errors was statistically significant (at $p < 0.05$), $t(11) = -2.88$, $p = 0.013$.

However, a higher error rate is not necessarily good evidence that participants are using social information. If they are, then we would expect that they are more likely to misclassify

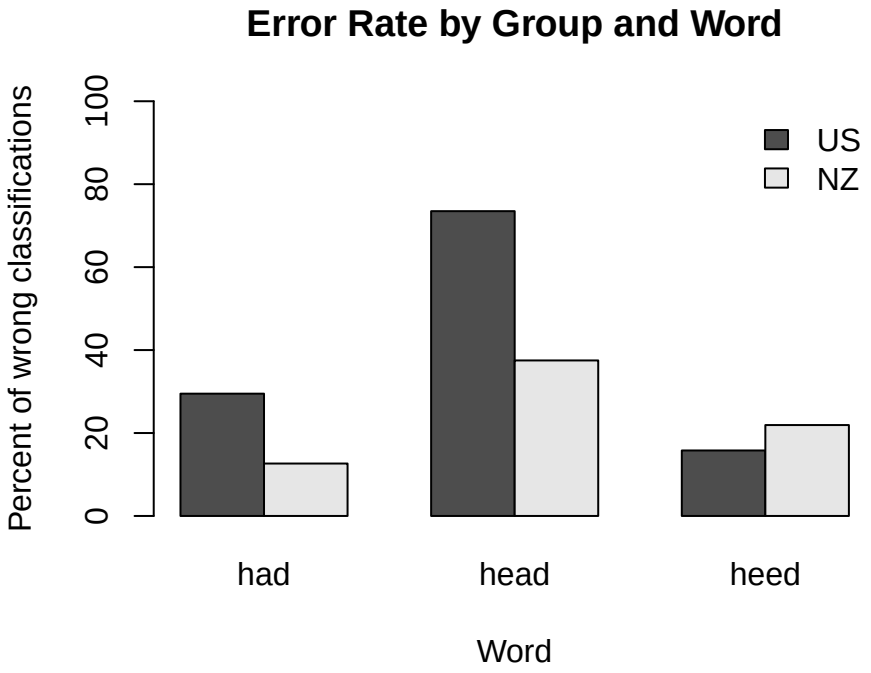

Figure 4.5: Error rate (as percent) by group and word. Note that, in line with the predictions outlined above, there is a higher error rate for "head" tokens in the US group, but a very similar error rate for "heed" across both groups.

tokens consistently as if they were US English. In particular, we'd expect tokens of NZE "head" to be more likely to be classified as tokens of US "hid". However, we would expect to see very similar classifications of "heed" as "heed", since the [i] vowel is very similar across these dialects. And this is in fact what we observe in the data, which is summarized in Figure 4.5.

We can see that the classification of "head" tokens as "hid" tokens is more common in the US group than the NZ group by comparing the confusion matrices for the two groups. These can be seen in Figure 4.6 on page 66. Note that, as predicted, the US group was very

likely to misclassify "head" tokens as "hid" ones[3]. In addition, they were also far more likely to classify "had" tokens as "head" tokens. This is unsurprising given the fact, as noted in section 4.2.1 above, the vowel in NZE "had" is quite raised. However, classification of "heed" was approximately equally accurate over both groups. Overall, the US group showed more errors than the NZ group, with kappa values of 0.48 and 0.66 respectively, and these errors are in keeping with the idea that participants in the US group are in fact treating the NZE vowels as if they were from a talkers of US English.

To summarize to results presented here: participants were all capable of learning to correctly classify NZE vowels with a relatively high rate of accuracy. When listening to a novel talker, however, only those listeners who were told that they were listening to another talker of NZE productively applied this perceptual learning. Participants who were told they were listening to a talker from the US classified the vowels they heard as if they had in fact been produced by a talker from the US. This suggests that listeners' social knowledge, even though in this case it was incorrect, was very important to their categorization of speech sounds.

4.3.1 Individual results

While the overall results show a clear pattern, the individual results show a surprising trend: only participants in the US group showed consistent improvement from the first to second task, as can be seen in Figure 4.7 on page 67. This improvement was statistically reliable for the US group (Wilcox signed rank test, $V = 0$, $p < 0.01$), but not for the NZ group (Wilcox signed rank test, $V = 10$, $p = 0.57$). While the latter is unsurprising–we'd expect no changes between the first and second task in the NZ group, since they were performing exactly

[3]One participant from the US group contacted the author to alert her that, though there were "hid" tokens in the second part of the experiment, they had mistakenly been left out of the training portion, which suggests a very strong "hid" percept!

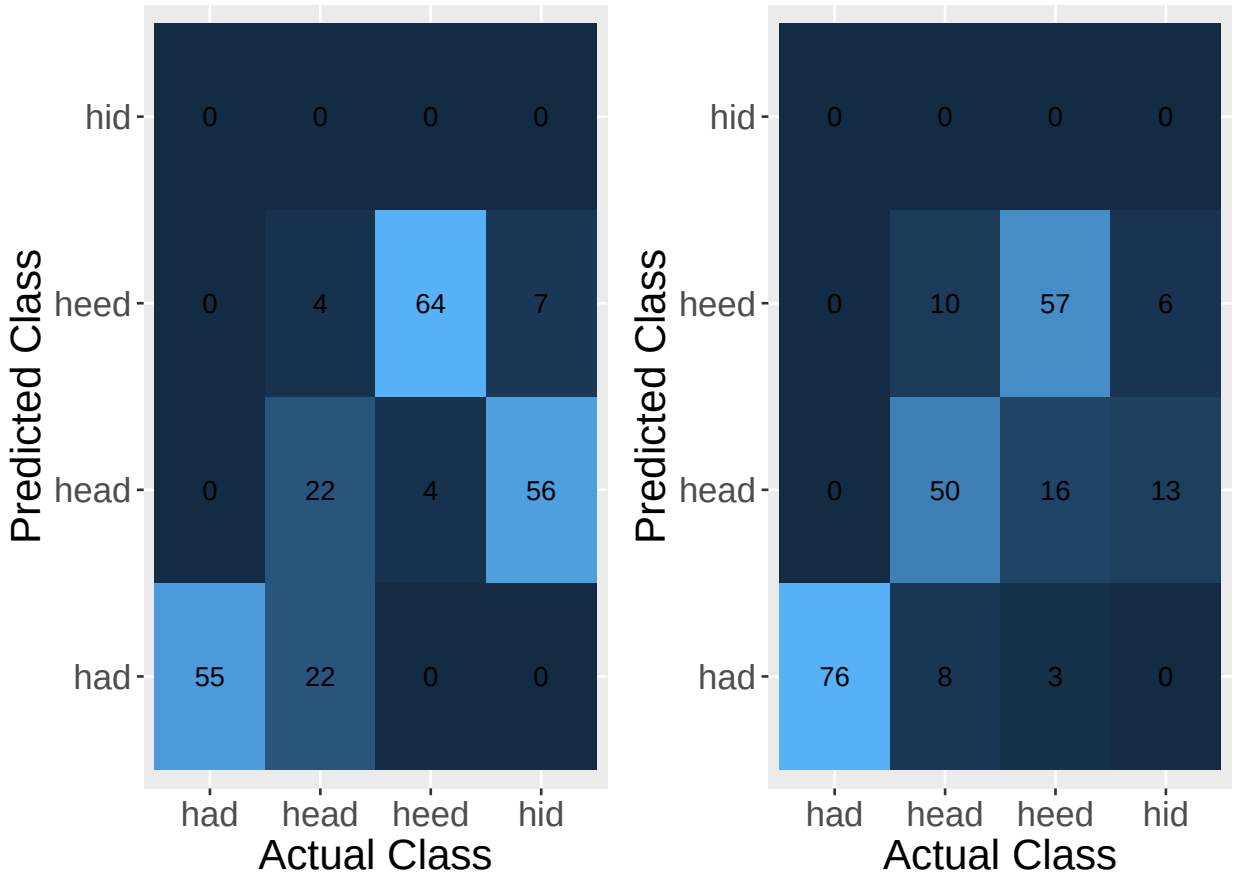

Figure 4.6: Confusion matrix for the NZ (left) and US (right) groups. Note that listeners in the US group were more likely to classify "head" tokens as "hid" than they were to classify them correctly.

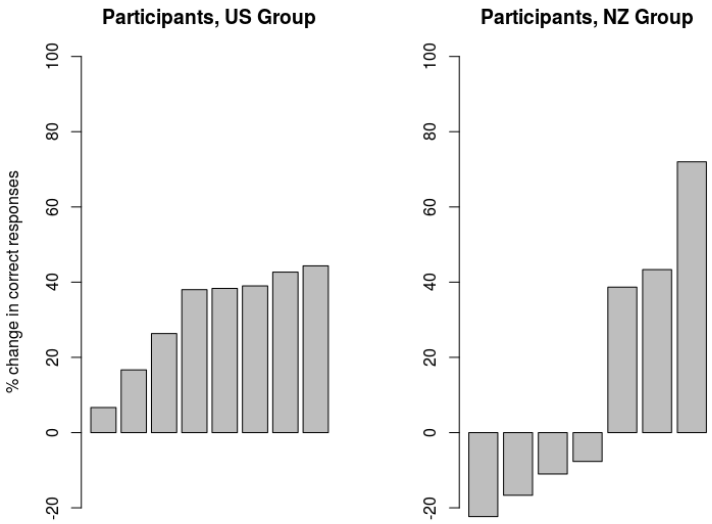

Figure 4.7: When looking at individual results, only participants in the US group showed a clear pattern of improvement between tasks. This pattern is all the more striking given that it includes these participants much lower performance on "head" tokens, as seen in Figure 4.6 on the previous page.

the same task and should have already reached titration–the former is very unintuitive. Especially given their poor performance on "head" tokens due to incorrect social information, we would expect these participants to fare worse.

Closer examination of these results by individual, however, reveal what is driving this improvement by the US group: better classification of "had" and "heed" tokens. This can clearly be seen in Figure 4.8 on the following page. This pattern can't be explained as only the effect of practice or growing familiarity with the task, or we would expect to see a similar effect on the NZ group–and there was none. That lack of effect in the NZE group is actually comforting; if participants were indeed reaching titration on NZE during the first stage of the experiment, then we should see no predictable improvement in the second half where they are repeating the same task.

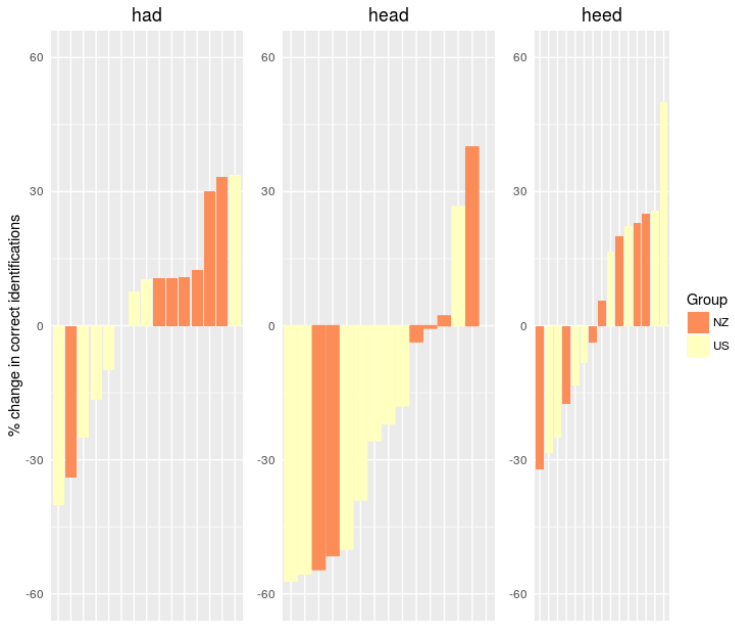

Figure 4.8: Changes in performance between the first and second task, by participant and item. Each bar represents a single participant, sorted from worst to best change between trails within each item.

The results for the US group, however, suggests that participants are doing two types of learning. The first is the expected perceptual adaptation, where participants learned to identify NZE vowels and then to correctly apply that learning. The second, however, is a more subtle type of perceptual adaptation that applies only when participants believe that the task has changed, and it is this that we see from the American English talkers.

This is not an entirely surprising result, though it was not anticipated at the outset of this experiment. There's a fairly strong body of evidence that learning can take the form of both top-down perceptual learning as well as bottom-up adaptation. One example from the latter is selective speech adaptation. In selective speech adaptation, participants shift perceptual phoneme boundaries after being repeatedly exposed to those phonemes (Cole, Cooper, Singer, & Allard, 1975). This effect is well attested (Bertelson, Vroomen, & De Gelder, 2003;

Dahan et al., 2008) and seems to be a process restricted to speech (Vroomen & Baart, 2009). A similar two-stage process, where one stage uses social information and the other focuses on fine phonetic detail, has been observed in dialect acquisition (Nycz, 2013). With this in mind, then, it is not shocking to see participants displaying two different types of learning.

However, based only on the current results, it is impossible to determine whether the distinction between the NZ and US group is actually based on whether or not they were applying social knowledge. In order to determine this, it is necessary to directly compare the same participants doing two tasks–one where social information suggests that they should be applying their perceptual learning and one where the social information suggests they should not apply it. It is in this second case where we should expect to see selective speech adaptation to the categories they have been exposed to during their training.

4.4 Conclusion

The results of this study provide evidence that:

- With training, listeners can learn to correctly categorize New Zealand vowels.
- This categorization can carry over to a second talker, however, it is conditioned by the given social information. If listeners believe they are listening to another American, they will categorize vowels consistently with American vowels, rather than with the New Zealand vowels they have just been extensively trained.

This is an important investigation because it will help us to understand how listeners make use of social information when learning a new dialect. Rather than passively shifting their linguistic system as a whole, they instead either maintain or abandon their newly-learned percepts based on the talker's identity. While this is unsurprising from a sociolinguistic perspective, it is more important for computational modeling. The fact that the persistence of

perceptual adaptation is talker (or at least social-group) specific suggests that, in some ways, talker-adaptation models of dialect adaptation that maintain separate models for different talkers (Gauvain & Lee, 1994; Leggetter & Woodland, 1995; Kuhn et al., 2000, and more recent work building on them) may be more cognitively accountable.

This aligns with findings that show that social information is closely tied to speech perception in general (Drager, 2010; Niedzielski, 1999) and may help to explain other studies where perceptual learning failed to carry over (Levi et al., 2011; Eisner & McQueen, 2005).

Chapter 5

MODELING LISTENERS' USE OF SOCIAL INFORMATION

5.1 Introduction

The experimental behavioral evidence outlined above suggests that social knowledge about a talker plays a large role in determining whether listeners choose to apply newly acquired perceptual learning. This is of interest both from a behavioral standpoint–linguists and cognitive scientists aim to describe human perceptual behavior accurately–but also from a modeling standpoint. Is it possible to construct a classifier which makes use of social information in the same way as human listeners? And can that model be similarly misled by incorrect social information?

5.2 Model

The ideal model for this application would have several qualities. First, it must be capable of multi-class classification. Secondly, it must be able to easily deal with features that are both continuous (formants) and categorical (dialect region). Finally, since this is behavioral model, it should be easily interpretable. Conditional inference trees, a specific type of decision tree, were chosen because they possessed all these qualities.

Decision trees are a model-class of branching tree structures, were each leaf is a classification or distribution over possible classifications. During classification, the classifier begins at the top node of tree. Each node evaluates a feature of the item to classified and then directs the classifier down one of the possible branches that emerge from it depending on the feature value.

For example, you can imagine a tree for English phonemic classification. The first node might test the [+/− syllabic] feature. Underneath the [+ syllabic] branch, the next node may test the [+/− consonantal] feature. Underneath the [+ consonantal] branch, the next node could test the [+/− voice] feature–however no such node would be needed from the [- consonantal] branch, since in English all vowels are voiced. (Note that if were attempting to classify a voiceless vowel, we'd be out of luck. Since we didn't put a node which evaluated voicing under the branch where we selected vowels, we won't be able to check for it during classification. As a result, voiced vowels will be misclassified as their voiced counterparts.) After continuing down the tree, evaluating a feature at each node, we would finally come to the terminal node of the tree, which would determine the classification of our data point–in this case the identity of our mystery phoneme.

5.2.1 Conditional inference trees

The models used here are conditional inference trees (Hothorn & Zeileis, 2014). Like many decision trees, they are trained by repeatedly splitting the feature space in two along a single feature, creating smaller and smaller areas until a predetermined degree of homogeneity is reached in all of them. In the case of conditional inference trees, that point is when no independent variable's explanatory power of the dependent variable reaches statistical significance. The advantage of this stopping method is that it is transparently established *a priori* and is directly related to the explanatory power of the variables under consideration, which is not necessarily true of other stopping methods, such as pruning a predetermined number of leaf nodes.

Each bisection of the decision space translates into one node of the resulting decision tree. If a large area of the decision region is particularly homogeneous, this area may be isolated with fewer cuts, and the branch may then terminate earlier than its sisters. It differs from

similar algorithms, however, in two ways. First, the selection of which feature to split on and the splitting itself is done in two separate steps and, second, the choice of which feature to split on is made through conditional inference rather than information gain.

In order to determine which feature to use in each node, multiple statistical significance tests are performed to select the feature which has the strongest significant relationship to the dependent variable. The method for selection of the most robust variable can be seen in Equation 5.1 (Hothorn & Zeileis, 2014). This allows us to calculate the distribution of case weights ($\boldsymbol{T_j(\mathcal{L}_n,}w)$), where X_j is a vector of the independent variable, Y is the dependent variable, w are the case weights, $\mathcal{L}_n$ is the learning sample and g_j is a non-random transformation of the co-variate X_j.

$$\boldsymbol{T_j(\mathcal{L}_n,}w) = vec(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i,(Y_1 ... Y_n))^\top) \in \mathbb{R}^{p_j q} \tag{5.1}$$

Note that features selected to split on this way may not be the more informative feature. For example, if we're trying to learn to predict an individual's occupation based on a selection of their traits, and it so happens that both the sanitation workers in our training set have blue eyes, this may be very informative (e.g. split the decision space in such a way that one group is completely homogeneous). However, eye color may not be a significant predictor of occupation in general in our training set. A model selecting for informativeness rather than significance may choose to use eye color as a feature for this node, while one based on significance probably wouldn't.

The conditional inference trees discussed here were produced using the ctree function in the Partykit R package, which builds upon the functionality of the Party R package (Hothorn, Hornik, & Zeileis, n.d.; Hothorn & Zeileis, 2014).

5.2.2 Features

One benefit of modeling human speech perception is the large body of literature there is to draw upon. As a result, feature selection can be done manually. For this study, the features used were static measures of the first and second formant (F1 and F2) taken at the midpoint of the vowel, as well as dialectal information, for a total of three features. Though static formant measures are not without their shortcomings (G. S. Morrison & Assmann, 2012), they are well-studied and widely used to capture differences in vowel quality. This makes them a good choice for an initial model, although later ones may use more complete formant models.

5.2.2.1 Feature scaling

Feature scaling is an important part of any machine learning project. It ensures that all features share similar ranges, which prevents the model from being misled by uninformative differences in the data. For example, you may be interested in modeling shoe size based on height in inches and income in dollars. If the range for the height measurements is from 60 to 75 inches and the range for our salaries is from 25,000 to 100,000 dollars, we're implicitly telling our model that a ten-dollar change in salary is as important as a ten-inch change in height–which is clearly not the case. By scaling these measures, for example from 0 to 1, we ensure that they're considered equally.

Vowels There are many popular vowel normalization methods in linguistics. For this particular application, it was necessary to find a method of vowel normalization which both reduces talker-specific differences and does not remove sociolinguistic differences. Fortunately, a direct comparison of common vowel normalization methods based on these criteria is already available (Adank, Smits, & Van Hout, 2004). Based on this analysis, the ear-

lier normalization method developed by Nearey (Nearey, 1978) was chosen. In addition to maintaining dialectal information, this method performed the best out of all the methods compared by Adank et al. on maintaining distinctions between the vowels included in this study. Normalization was done using the norm.nearey function from the Vowels R package (Kendall & Thomas, 2009), which is also available though a web front end: NORM[1].

Dialect information

5.2.3 Effect of participant dialect

One key benefit of the Conditional Inference Tree implementation used here is that it can handle categorical data without the need to convert it to a numeric score. Statistical tests of categorical data are done through creating a vectorized contingency table and submitting it to a chi-square test.

5.2.4 Training

All models discussed here were trained in the same way, with the classification output being the word ("had", "head", "heed" ,"hid") and the features used for classification being talker dialect and Nearey-normalized F1 and F2. Models were trained and evaluated using sub-sampling cross-validation, with half of the data randomly-selected for training and the rest used for validation.

5.2.4.1 American English data

In order to model perceptual adaptation, it was necessary to include speech data taken from a non-NZE talker in the training data. Because the participants in the experiment were

[1]http://ncslaap.lib.ncsu.edu/tools/norm/

American English talkers and presumably had the greatest exposure to that population, speech data from an American English talker was used. An American English talker was chosen to who matched the sociological characteristics of the NZE talkers–a young, white, university-educated woman[2]. Tokens for the US talker were recorded in a sound-attenuated booth while wearing an AKG Pro Audio C520 L Head-Worn Condenser Microphone. The recording was taken using Audacity (Mazzoni & Dannenberg, 2000) with a sampling rate of 44.1 kHz. 17 tokens each for "heed", "hid", "head" and "had" were recorded in isolation.

5.3 Results

Two conditional inference trees were trained for this task. Both classifiers were trained in the same way, with word as the classification output and Neary-normalized F1 and F2 and, in the second model, speaker dialect as features. Models were trained and evaluated using sub-sampling cross-validation, with half of the data randomly selected for training and the rest used for validation. Both models achieved the same classification results on the original data with the correct dialect information, with a balanced accuracy of 0.86.

The second model, which included dialect information, is summarized graphically in Figure 5.1. The thing to take especial note of here is that the classifier, when given the option, *does* depend on the dialect of the talker. This is very heartening, because it mirrors the performance of the human participants, who also used dialect information in their classification. This can be further verified by comparing confusion matrices from the human participants and classifier, which are shown in Figure 5.2. (Note that the classifier did have the advantage of positive examples of "hid" from both dialects. This was possible because duration was not given to the classifier, so there was no worry about duration information confounding classification.) However, there is one disheartening wrinkle–a model with only F1 and F2 as

[2]Since the social stratification of the US and New Zealand is somewhat different, university education was used as a proxy for social class.

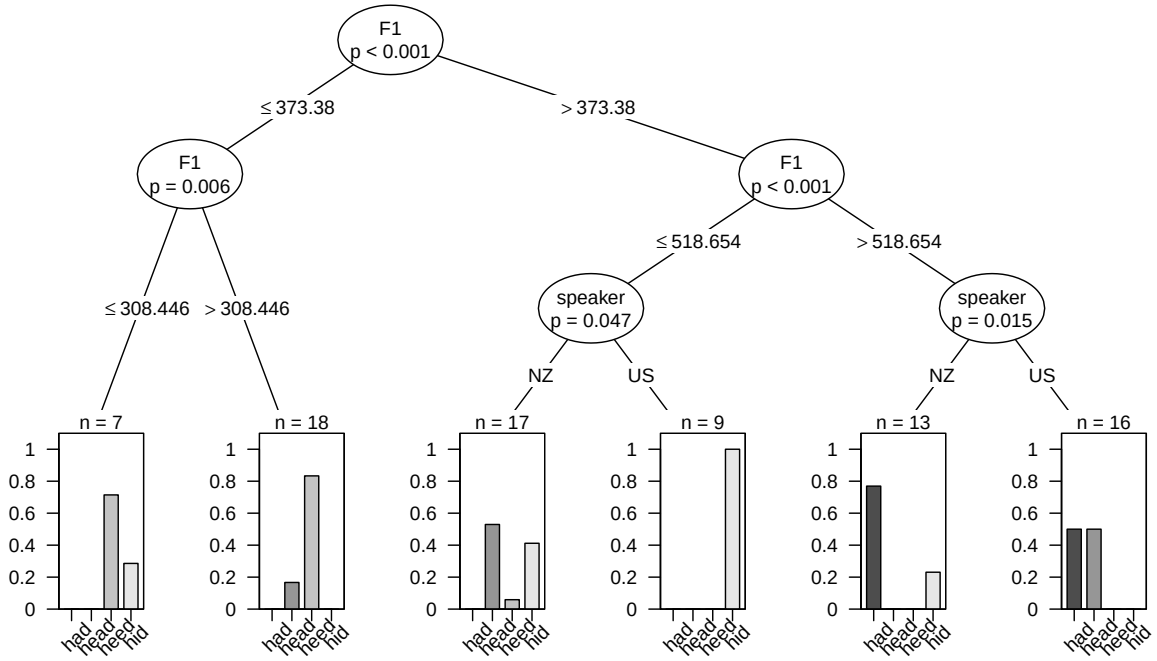

Figure 5.1: Graphical representation of decision tree trained on data. Note that two different nodes use the "speaker" feature, which contains the dialect of the talker.

input not only has the same accuracy (96.5%), but also precisely the same classifications. At first blush, it looks like including social information about talkers is an extra step that does nothing to improve classification. This impression changes, however, when these models are evaluated to see if they make the same sort of mistakes that human participants did if it is given incorrect information about a talker's dialect region.

To determine this, the test data set was all assigned either "US" or "NZ" as dialect. While this is not entirely parallel to the human participants–who were only given mis-labeled American English and correctly labeled NZE–it is a good proxy. It also allows us to make a behavioral prediction. If this model is capturing a generalization about human behavior, then it should make a prediction about how human listeners will react when presented with American English data, whether correctly or incorrectly labeled.

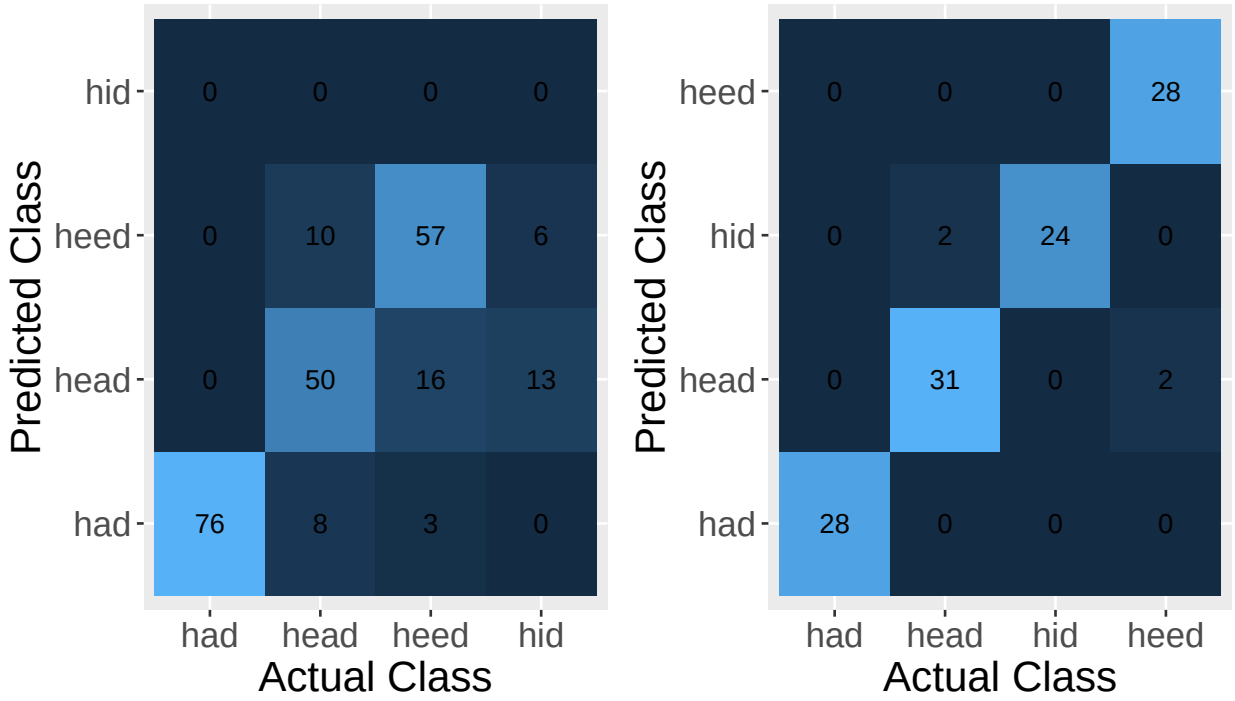

Figure 5.2: Confusion matrix for human (on right) and conditional classification tree (on left). The automatic classifier does perform better than humans in this instance, but note that the data here is taken from the entire training session. The high proportion of errors can partially be attributed to mistakes made during learning.

Figure 5.3 shows the true and automatic classifications of both the US and NZ data. This figure is especially exciting because it shows how the classifications change when presented with incorrect information about the dialect of the talker. Note that when the dialectal information is incorrect, tokens are classified as the most similar tokens in the other dialect, which is precisely the same behavior as the behavioral participants displayed–at least for data that was mislabeled as being from the US, as shown on the bottom left. The classification on the bottom right, however, has not yet been experimentally verified.

On the other hand, Figure 5.4 shows that the model without talker information will not change its classifications when that information is changed. This is entirely unsurprising. This figure is included here only to highlight the difference between this model and the previous one. It is also something of a warning. While Occam's Razor would suggest that, given that these two classifiers produced the same classifications, the better one would be the second, as it has fewer features. (Never mind the difficulty of getting accurate and useful social information when working at a larger scale.) However, if the goal is human-like behavior, as it is in behavioral modeling, it is important to include features that may not improve the initial classification if there is strong evidence that listeners are using those features.

The classification discussed here have shown that dialectal information can be usefully applied in classification of speech sounds, and that incorrect dialect information can lead to classification errors for confusable sounds. In addition, the model described here makes a behavioral prediction–human participants should misclassify American English sounds mislabeled as NZE.

The model which includes talker information is not, however, a perfect proxy for participant behavior. While it does succeed in using dialect information in an intuitive way, it does

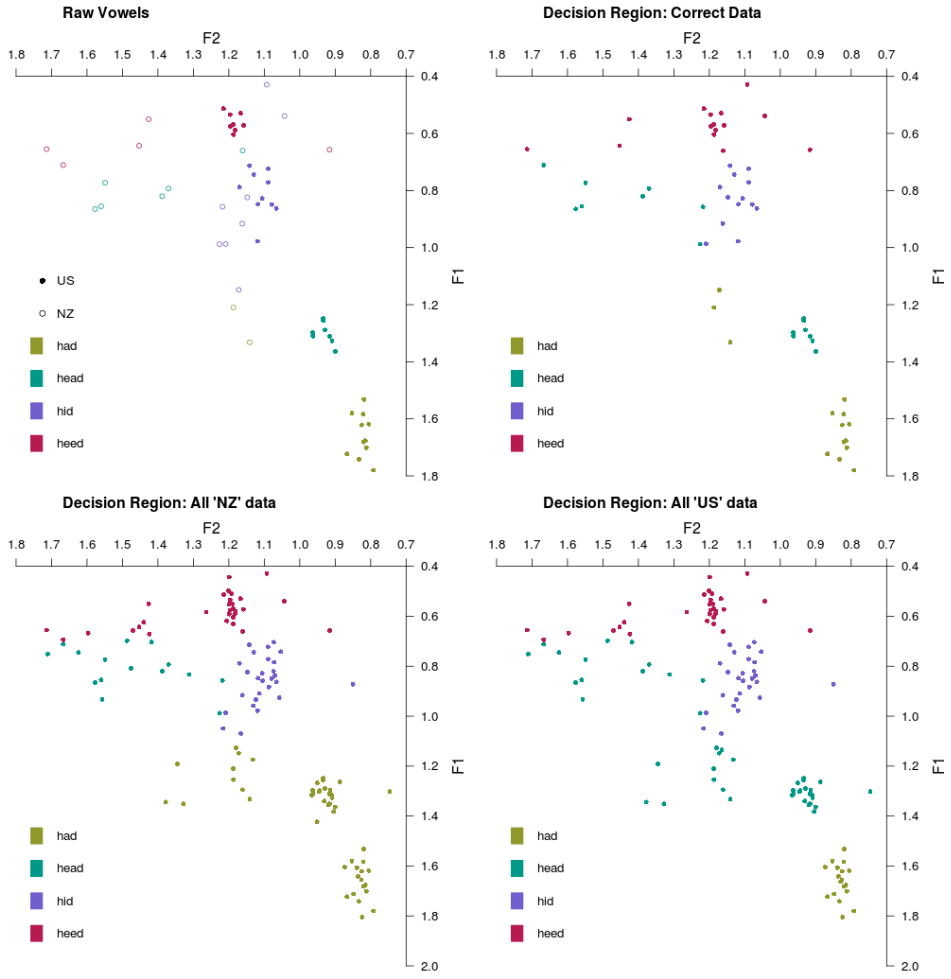

Figure 5.3: Correct and automatic classifications of the data set. Note that in the two bottom figures, the dialect for all items was changed to either "NZ" , on the right, or "US", on the left.

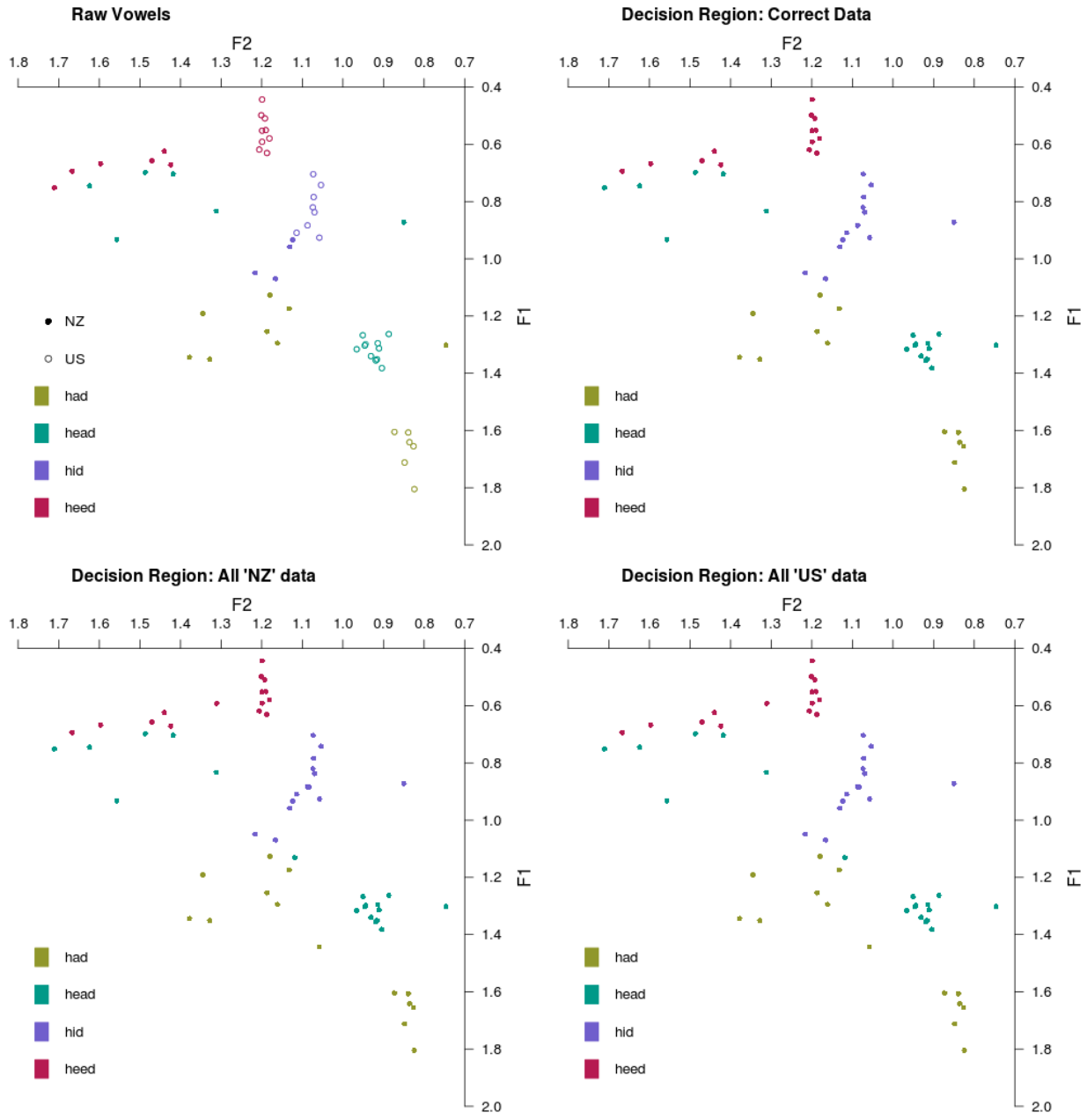

Figure 5.4: Decision areas for the model trained without talker data. Note that–not surprisingly–this results in a model which does not change its classifications based on social information about the talker. So while including talker information does not, in this case, improve classification accuracy, it does make performance more human-like.

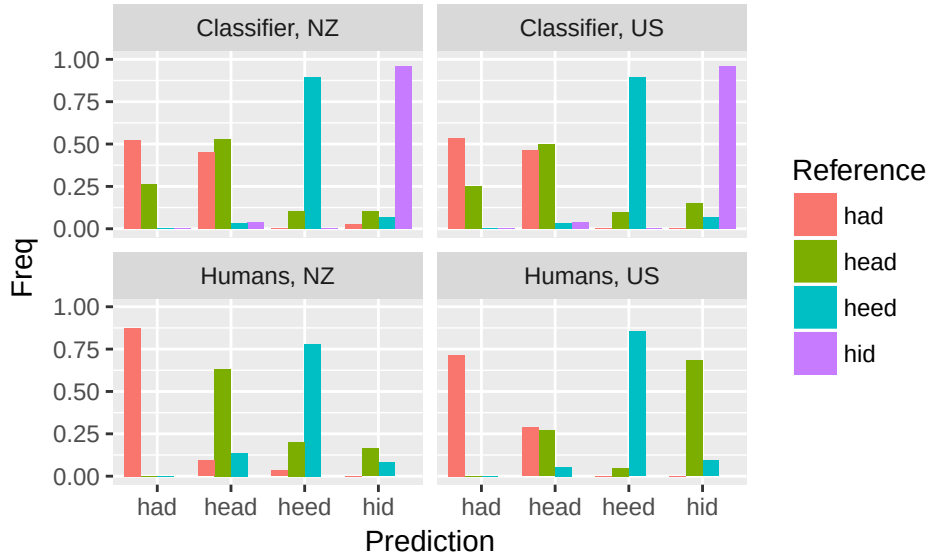

Figure 5.5: Classifications by class for participants and classifier on data all labeled both "US" and "NZ"

not make precisely the same type of classification errors as the human participants. This can be clearly seen when looking at the proportion of classifications shown in Figure 5.5. Note that these errors differ between the automatic and human classifications. Of particular interest is the fact that the classifier did not classify "head" items as "hid" items when told that NZ items were produced by a US talker. In fact, as can be corroborated by looking at all the figures in this section–the classifier almost always correctly classifies "hid" as "hid", regardless of the dialect of the talker. This deviatesf from the participants' classifications errors.

There are several possible reasons for this deviance. First, it may be an artifact of the training data. Since data was taken from only one American English talker, it has a much higher degree of internal consistency than the data for the NZE talkers. Secondly, the data set used here is quite small. There were only 45 total tokens of New Zealand English across all vowels, which is minuscule training set in the Automatic Speech Recognition (ASR)

literature, where it is far more common to see hundreds of hours of speech used to train classifiers. Both of these shortcomings may be overcome with the careful addition of more data. A third, and more concerning possibility, is that feature selection for this model has resulted in the loss of key information. Since it is already established that cues other than static formant measures are important in distinguishing NZE vowels, it is very possible that participants are using some of these features in their classification. The next stage of modeling should thus include both larger, more varied data and a more nuanced feature selection process. Considering the small size of the training data, however, the current results are quite promising, especially the behavioral predication about how participants may react to American English speech mislabeled as NZE.

5.4 Behavioral predictions and validation: Reaction times

In addition to predictions about the general classification behavior of participants, the conditional inference tree presented in Figure 5.1 also makes another prediction. Though I make no claims that it is a cognitive model, if this conditional inference tree is accurately capturing behavior then it suggests that listeners are processing "heed" tokens differently from "had" and "head" tokens. The two leftmost leaf nodes capture almost all "heed" classifications. These two terminal nodes also share two other key qualities. First, the path from the root node to these terminal nodes is shorter than the paths to the other terminal nodes. Second, it is not necessary to pass through any nodes which split based on talker identity to reach them. This captures the fact that "heed" tokens in NZE and MUSE are fairly similar and tightly clustered. It follows, then, that if classification of "heed" requires less fine-grained distinctions and also that it does not require accessing recent dialect-based learning, then classifications of "heed" should be faster than classifications of "head" or "had".

And, in looking at the behavioral evidence, we find support of this. Classification reac-

Figure 5.6: Classification reaction times for "heed", "hid" and "had" tokens from the application task. Note that, in keeping with the predictions outlined above, classifications of "heed" are significantly faster than classifications of "had" or "head".

tions times for "heed" tokens during the application phase were faster than for other tokens, as can be see in Figure 5.6. The median reaction time for "heed" tokens was 1000 ms (or one second), while the median for "had" and "head" was 1076 ms and 1198 ms respectively. (The reaction times over all were fairly slow, but that is to be expected given both the off-line nature of the task and latency due to render time and internet connections.)

In order to verify that these differences were robust, a linear mixed effects regression for reaction time was constructed that included subject and whether the response was correct as random effects (slopes and intercepts). A model which also included item as a fixed effect

performed significantly better than one that did not ($\chi^2(2, 635) = 24.7$, $p < 0.01$).

The experimental reaction time data closely patterns with the general pattern captured by the conditional inference tree: that "heed" is easier to classify across these two dialects than "head".

5.5 Conclusions and further work

The experimental part of this study presents strong evidence that, after successful perceptual learning of a vowel system that differs from their own, participants then choose whether to apply that perceptual learning to a new talker or not based social information about that talker. This joins a robust body of literature that points to the importance of using social information during speech perception, and may help to resolve some puzzling instances where participants failed to apply perceptual learning in a new situation.

The modeling part of this study presents evidence that this same social information can effectively be incorporated during machine learning. This is not surprising in light of the large number of classifiers which have also successfully improved accuracy by applying dialect information during ASR (Humphries et al., 1996; Lincoln et al., 1998; Huang & Hansen, 2007; Aubanel & Nguyen, 2010; William et al., 2013). While the introduction of social information did not, in this case, improve overall classification accuracy, it did improve how closely the model matched the participants' behavior when presented with incorrect feature values. In addition, it offers a testable behavioral prediction–participants presented with mislabeled data from American English talkers should classify them as if they were NZE as well. A second behavioral prediction, that "heed" tokens should be classified faster than other tokens, was empirically verified using reaction time data.

Further work will focus on three main trajectories. The first is to verify the behavioral prediction made by the model. It is possible–although very unlikely–that whether perceptual

learning carries over or not is simply a matter of individual differences, and that the random assignment of participants to the US or NZ groups resulted in a separation between these groups. In order to ensure that this is not the case, a second experiment including correctly and incorrectly speech from multiple dialects is necessary.

The second main line of inquiry will be in refining the modeling of the use of social information. This will be done through the use of larger data sets, including incorporating behavioral data from the second set of experiments, as well as the use of more nuanced features. In particular, duration and vowel inherent spectral change may be key perceptual features that should be the focus of additional work. Additional work may also look at automatic feature extraction for the dialect feature. This is an area which has received more attention in the computational literature and including it will improve the extensibility of the new models.

The third main line of inquiry will be to investigate the role of social information during perceptual learning in other contexts. Can participants be trained concurrently in multiple social distinctions? Can these findings be extended to non-vowel speech sounds? Are social distinctions other than dialect region considered similarly? On an even broader level, does this extend to other languages? What is the difference in use of social information between, say, dialects of English as opposed to English and Frisian?

The work outlined above here is clear evidence that, if human-like behavioral is desirable during Automatic Speech Recognition, then social information must be included both during training and classification. Without the relevant social information, automatic classification does not display the same general types of errors as human listeners and, at a larger scale, may produce less accurate classifications overall. This may be especially relevant when working with sparse training sets, such as for under-documented languages. If the data available comes from multiple dialects, including dialect information during training should result in

more human-like recognition.

While the scope of the current study is necessarily narrow, further work on this topic will hopefully build upon it and result in a simple, accountable way of including the social information that we know humans use in machine learning of speech sounds.

Chapter 6

THE IMPACT OF SOCIAL INFORMATION DEPENDS ON LISTENERS' FAMILIARITY WITH THE DIALECT THEY ARE HEARING

Experiment one and the model derived to approximate it, described above, provide evidence that listeners choose whether or not to apply recent perceptual learning based on information about where the talker is from. However, the study design had participants do both training and testing on talkers from another dialect. This leaves the possibility open that listeners only rely on social information when listening to a dialect they are unfamiliar with. While the speech samples given to participants were very short–only 150ms from the center of the target vowel–there is evidence that dialectal information is encoded in the vowel dynamics. A second experiment is necessary in order to determine the relative importance of top-down and bottom-up cues to talker identity. The former is the information given to participants about talker identity, and the latter the dialect cues embedded in the formant dynamics of the vowels themselves.

6.0.1 Formant dynamics of NZE and MUSE vowels used in this study

Because this experiment uses natural speech tokens rather than resynthesized ones, there was limited control over the vowel dynamics–that is, changes in the formants of vowels over the time course of the vowel. As can be seen in Figure 6.1, there were differences in vowel dynamics between these two dialects in addition to differences in formant values at the vowel midpoints. If the dynamics were static and only the formant values differed between these

two dialects, we'd expect to see parallel formants across the duration of the vowel. Figure 6.1 clearly illustrates that that is not the case.

These differences are not surprising given the growing body of evidence that formant dynamics, also known as Vowel Inherent Spectral Change (VISC) are dialect-dependent (G. S. Morrison, 2013). Koops, for instance, found that while Anglos in Houston all fronted their /u/, younger talkers had a more diphthongal fronted /u/, consistent with the West, while older talkers had a more typically Southern monophthongal fronted /u/ (Koops, 2010). There are similar differences between NZE and MUSE English with regards to VISC, especially for the "head" vowel, which is strongly diphthongal in this NZE data. As a result, it is possible that listeners will be able to use formant dynamics as a bottom-up cue to identify which dialect the tokens are from.

6.1 Method

The paradigm was very similar to the one depicted in Figure 4.2, with some modifications.

Before beginning the experiment, listeners took a demographic questionnaire to ensure that they were from the United States, that they had not traveled to New Zealand and that they were unfamiliar with New Zealand English. Any listener not meeting these criteria was excluded from the experiment.

Listeners completed three lexical decision tasks. In each task, they were played a 300ms audio sample taken from the steady-state of the vowel from one of the words "heed", "head" or "had". Only the vowel was used in order to avoid possible confounds of variation in stop-production across dialects (Lyle, 2008). As before, "hid" tokens were excluded due to the very salient duration contrast in NZE between the vowel in "hid" and other vowels (Langstrof, 2009). Audio data was taken from two talkers of NZE and one talker of MUSE. All talkers were white women between the ages of 20 and 30, with a college-level education

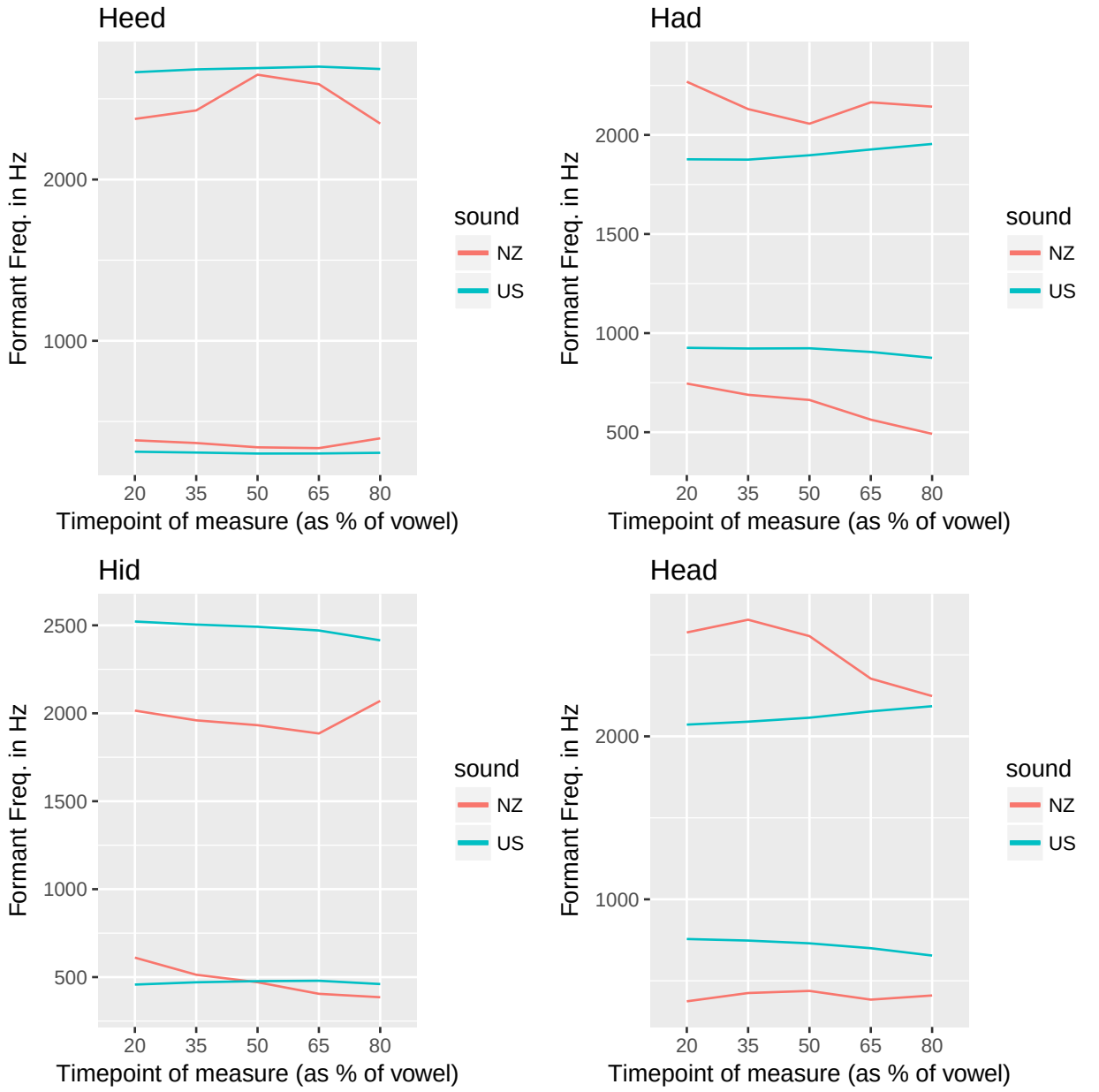

Figure 6.1: Plots comparing the formant dynamics of New Zealand and American English vowels. The central line shows the mean format measure at each time point for each dialect, while the shaded area is the 90% confidence interval. As can be seen in these charts, differences between NZE and MUSE vowels include robust differences in vowel dynamics as well as differences in formants at midpoint of the vowels.

or higher.

The first lexical decision task was designed to train listeners to correctly identify the vowels of NZE. Listeners were played a vowel and asked to input which word it had come from: "heed", "head", "had" or "hid". If they picked the right word, they were told that it was correct and played a different token. If they picked the wrong one, they were told which word the token had been taken from and were given another chance to answer correctly. This process continued until each listener correctly labelled ten tokens in a row. For this task, listeners heard audio from a single talker of NZE and were told that the talker was from New Zealand.

For the next two tasks, listeners did not receive feedback. In one task they were played audio from a MUSE talker, and in the other from a NZE talker. They were explicitly told before each task the national origin of the talker they were about to hear: New Zealand or the United States. However, while half of the listeners were given the correct information about each talker, the other half were given incorrect information about each.

The complete experimental paradigm for this experiment can be seen in Figure 6.2.

6.1.1 Participants

Twenty one listeners participated in this experiment, most of whom (17) were from the West, a US dialect region which stretches from the Rockies to the Pacific Ocean (Labov, Ash, & Boberg, 2005). The experiment was administered on-line using PsyToolkit (Stoet, 2010). Participants were recruited using social media and word of mouth.

6.2 Results

All participants were able to learn to successfully label the vowels of New Zealand English. This can be seen be seen in Figure 6.3: when given the correct social information about the

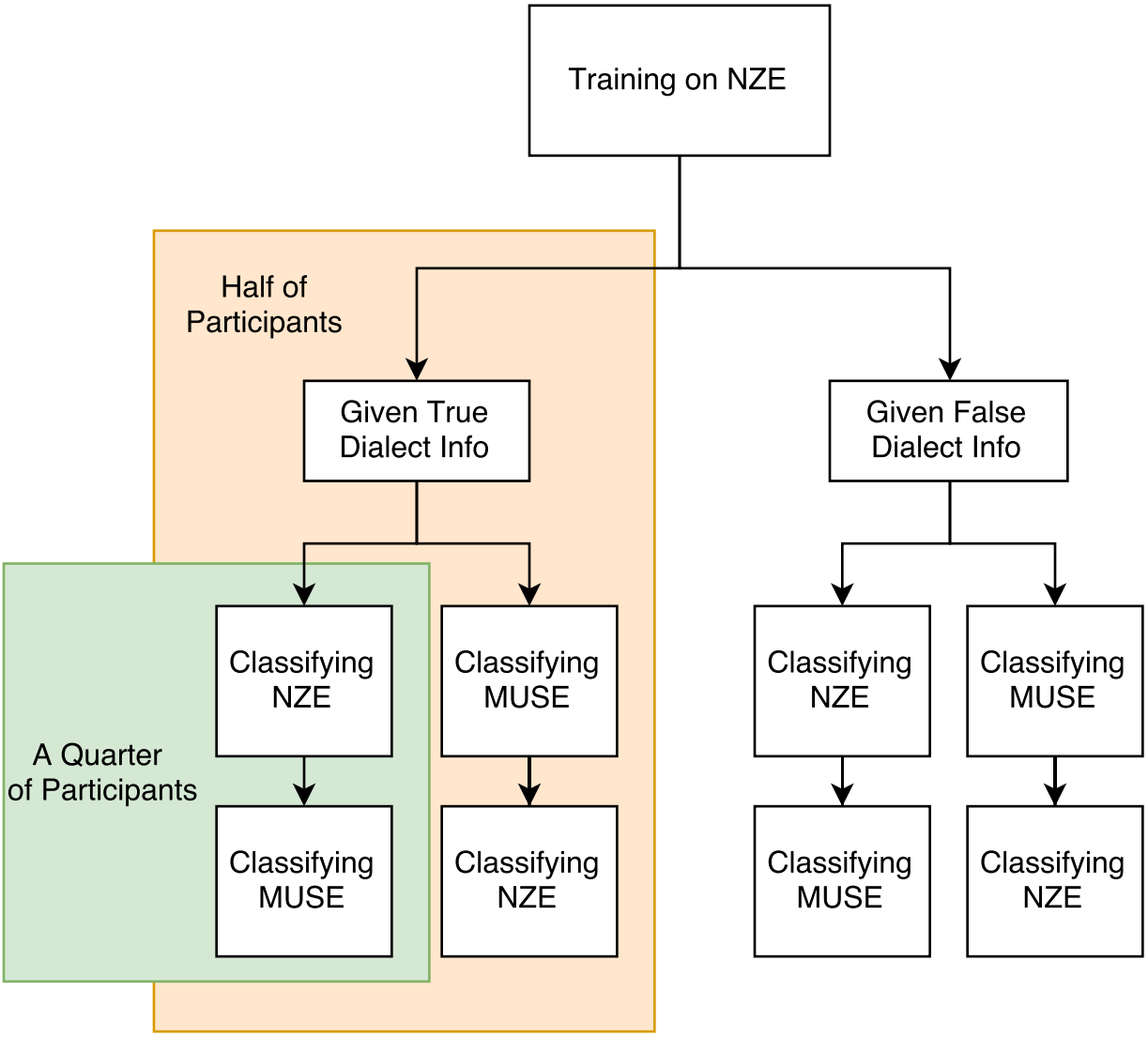

Figure 6.2: Experimental paradigm for the second experiment.

talker, the classifications of vowels from both MUSE and NZE were overwhelmingly correct, with an average F_1 across classes of 0.96 and 0.75, respectively.

Being given incorrect information about a talker's national origin, e.g. being told that a talker from New Zealand was from the United States affected listeners classifications differently depending on the actual regional origin of the talker. While listeners were still fairly accurate on their own dialect ($F_1 = 0.81$), they categorically changed their judgments for NZE ($F_1 = 0.56$). This difference can clearly be seen in Figure 6.4. The categorical changes are in line with confusions discussed previously. NZE "head" is mis-classified as "hid", and NZE "had" is mis-classified as "head".

In other words, when listeners from the United States are listening to a talker from New Zealand that they have been told is also from the United States, they are classifying NZE vowels as if they were from MUSE. However, the inverse is not true. Even when told that a MUSE talker is from New Zealand, listeners are still classifying their vowels as if they believed them to be from the United States.

6.2.1 Reaction time data

As with the first experiment, reaction time was collected for each trial. As in Experiment 1, there was a significant effect of item on reaction time. The addition of a second dialect, and both correct and incorrect participant assumptions about it, however, presents additional complexity.

6.2.1.1 Reaction time by item

As in Experiment 1, reaction times were fastest for "heed" and tokens and slowest for "head" tokens. This is summarized in Figure 6.5. The mean response time for correctly labeled "heed" tokens was 1119 ms, "had" 1162 ms and "head" 1340 ms. Once again, a linear

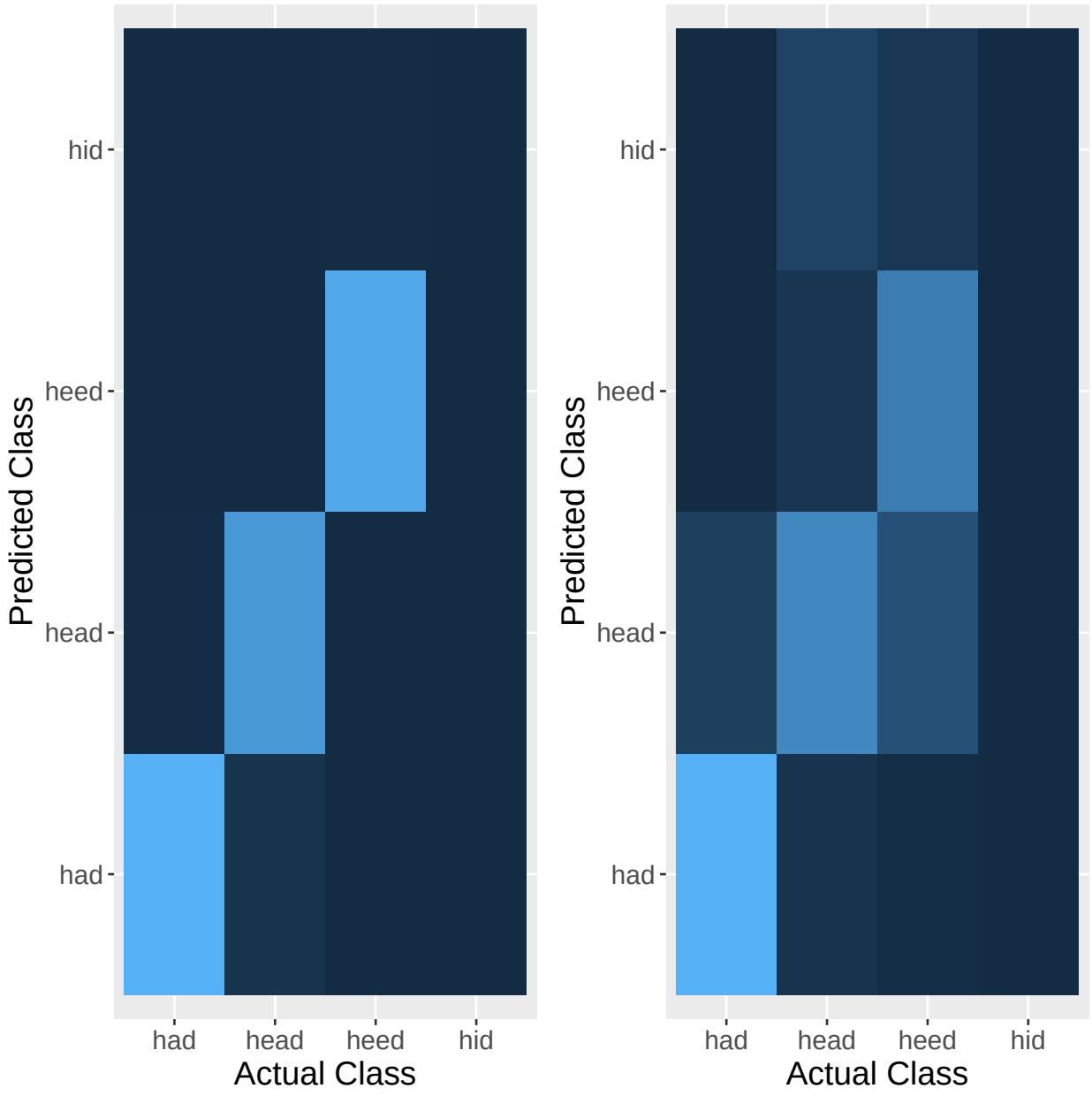

Figure 6.3: Heatmaps of the confusion matrix for MUSE (left) and NZE (right) classifications when participants were told the correct regional origin of the talker. Note that, in both cases, most classifications were correct, i.e. on the center diagonal.

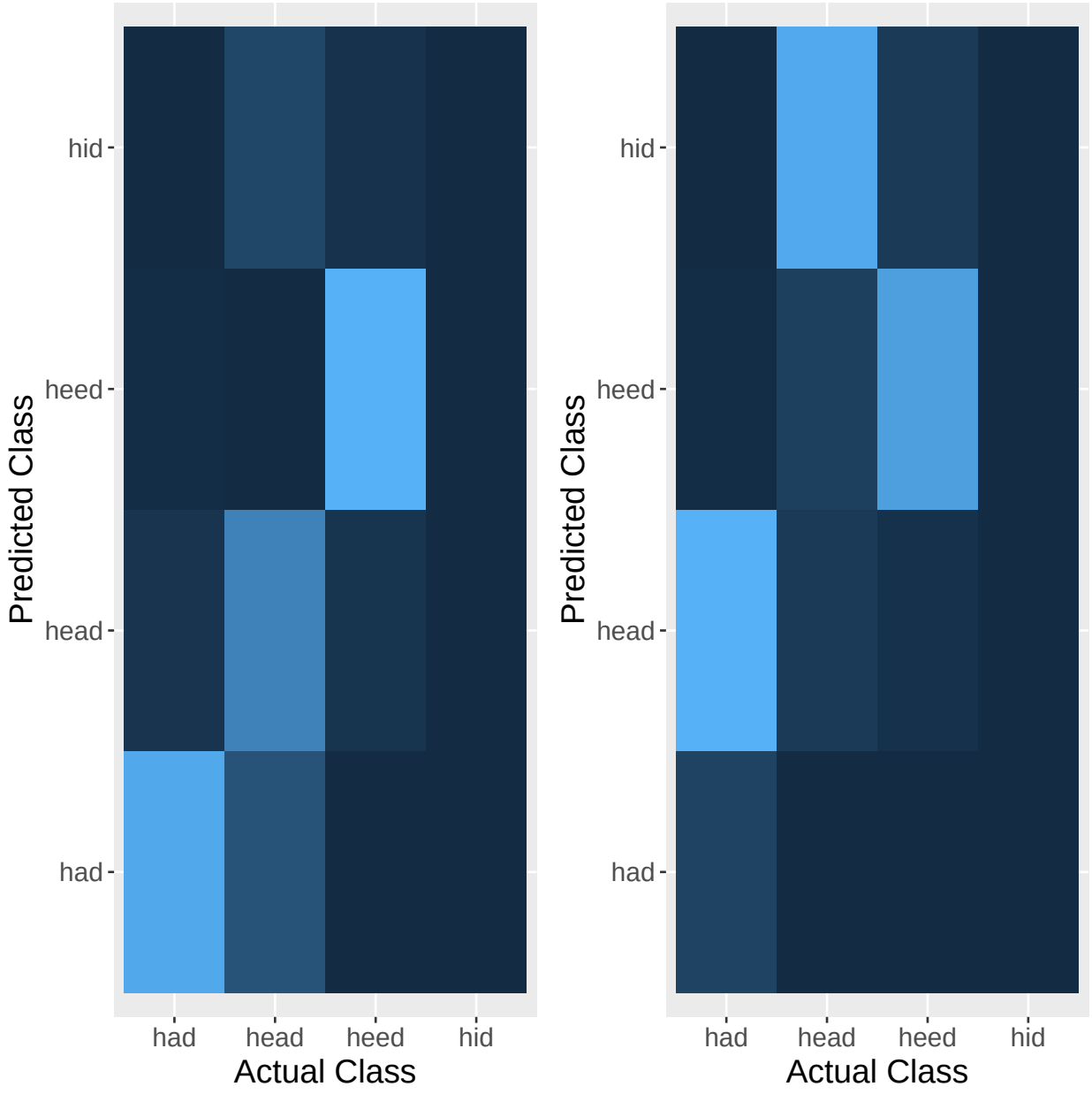

Figure 6.4: Heatmaps for the confusion matrix for MUSE (left) and NZE (right) classifications when participants were told the incorrect national origin of the talker. While this had a slight effect on the classification of MUSE vowels, it categorically changed that of NZE vowels--this is what is meant by "own-dialect bias".

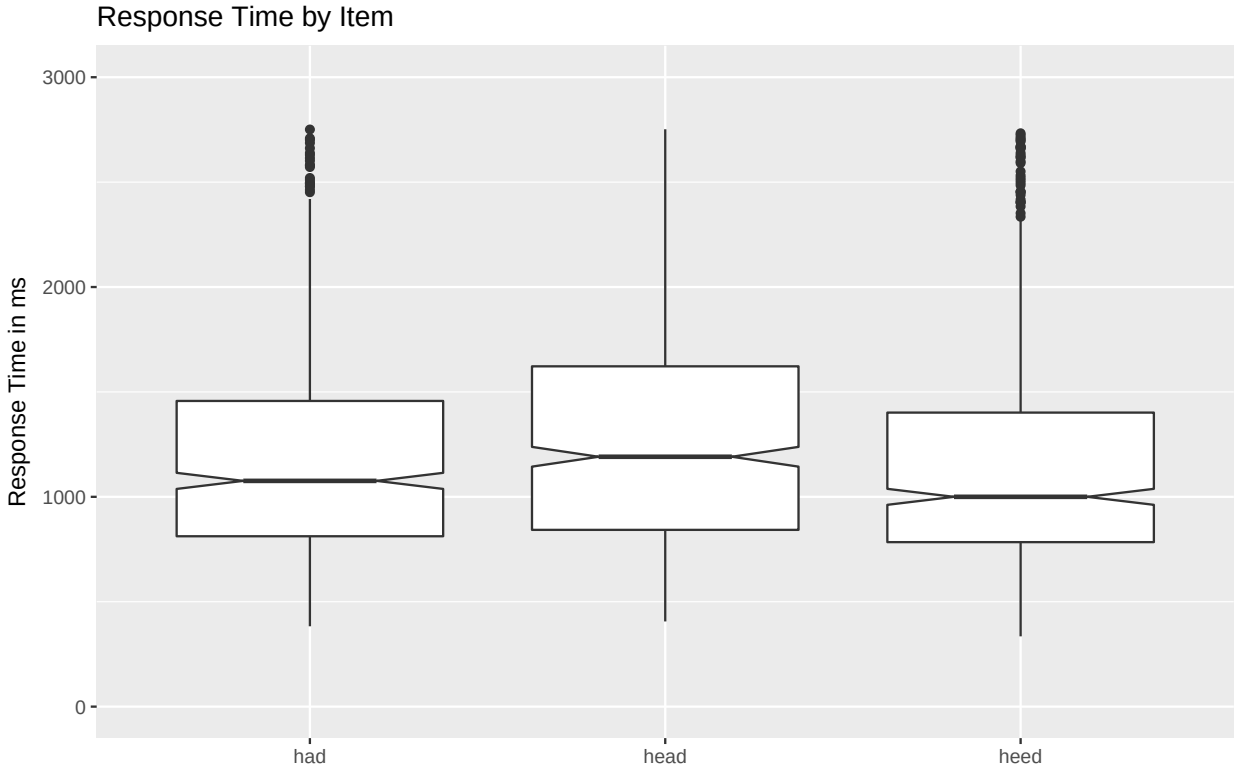

Figure 6.5: Reaction time for tokens, separated by vowel quality. As in Experiment 1, reaction times were fastest for "heed" tokens.

mixed effects model with subject and whether a response was correct as random effects was significantly improved by the addition of vowel quality ($\chi^2(2, 4029) = 28.35$, $p < 0.01$).

6.2.1.2 Reaction time by dialect & label

The dialect from which the token was drawn also had an effect on participant reaction time. The average reaction time for tokens from MUSE was 1144 ms, compared to 1277 ms for NZE tokens. Adding dialect to the maximal mixed effect model from Section 6.2.1.1 significantly improved model fit ($\chi^2(2, 4029) = 12.79$, $p < 0.01$).

Including whether the labeled dialect was correct or not, however, had no significant improvement on model fit ($\chi^2(2, 4029) = 0.46$, $p = 0.49$): there was no robust effect of

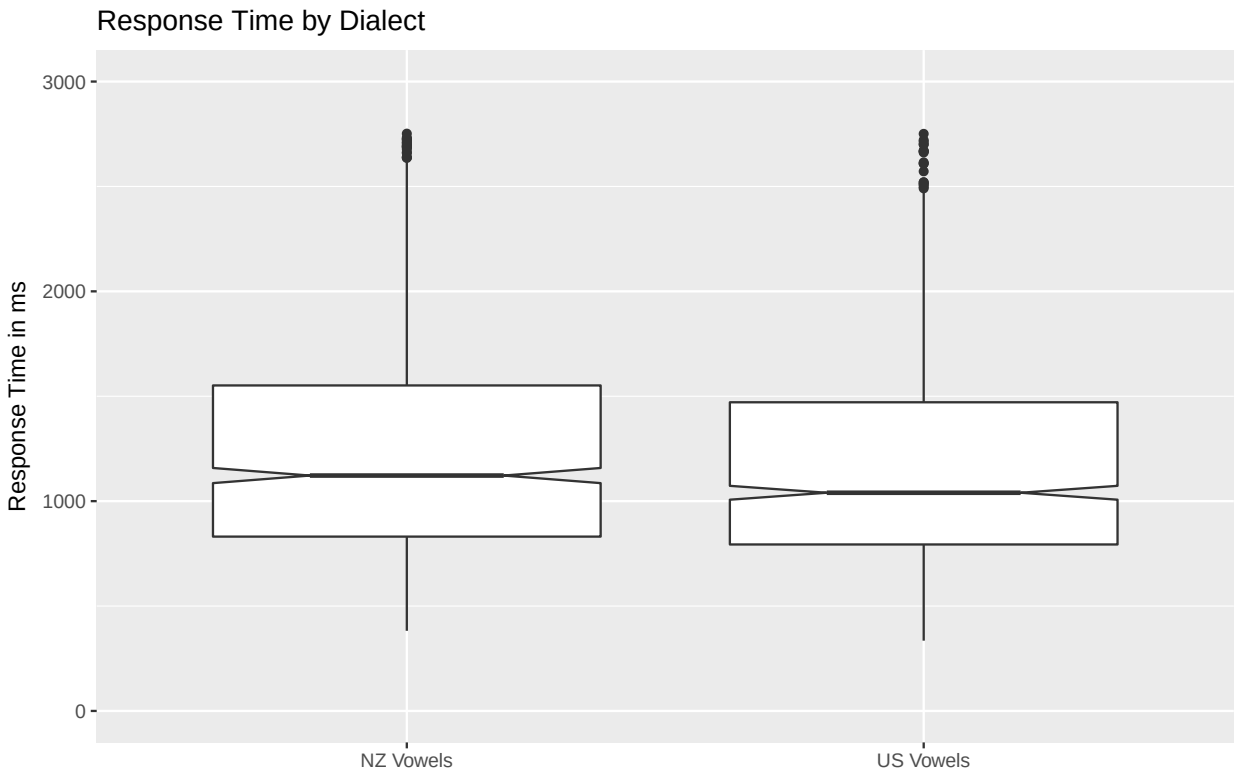

Figure 6.6: Reaction time by dialect. Reaction time was significantly faster for tokens from MUSE, the participant's own dialect.

whether participants were being lied to on the speed of their reactions.

There was, however, a strong interaction between the actual dialect origin of a token and whether it was correctly labeled. Tokens which participants believed to be from MUSE were labeled more quickly than those they believed to be from NZE. This effect can be clearly seen in Figure 6.7. Including the interaction between the item dialect and its label to the maximal linear mixed effect model described above significantly increased its fit ($\chi^2(2, 4029) = 153.65$, $p < 0.01$).

Figure 6.8 shows the interaction of item, labeled dialect and the actual dialect and is a good summary of the effects captured by the maximal linear mixed effects model described above.

To summarize: both the actual dialect source of a token and participants' beliefs about the dialect source of a token affected reaction times, but not the accuracy of their beliefs. Participants responded more quickly if a token was from their own dialect or if they believed that it was. This, along with participant responses, provides strong evidence that listeners are relying on both acoustic cues and social beliefs about talker's identities during speech perception.

6.2.1.3 Which is more important, a talker's dialect or listener's belief about a talker's dialect?

Above, I provided evidence that both a talker's beliefs about what dialect they are listening to and the dialect they are actually listening to affect response times. Which is more influential, though? In order to investigate this, a new mixed effects model was constructed which, rather than the interaction between dialect label and its truth, included which dialect listeners believed they were listening to. This model included subject and whether a response was correct as random effects, and item, actual dialect and believed dialect as fixed effects.

```
Data: appDatRT
Models:
baseModel: rt ~ item + block + (1 | subj) + (1 | corr)
advancedModel: rt ~ item + block + truth * block + (1 | subj) + (1 | corr)
              Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
baseModel      7 31073 31112 -15529    31059
advancedModel  9 31002 31053 -15492    30984 74.134      2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

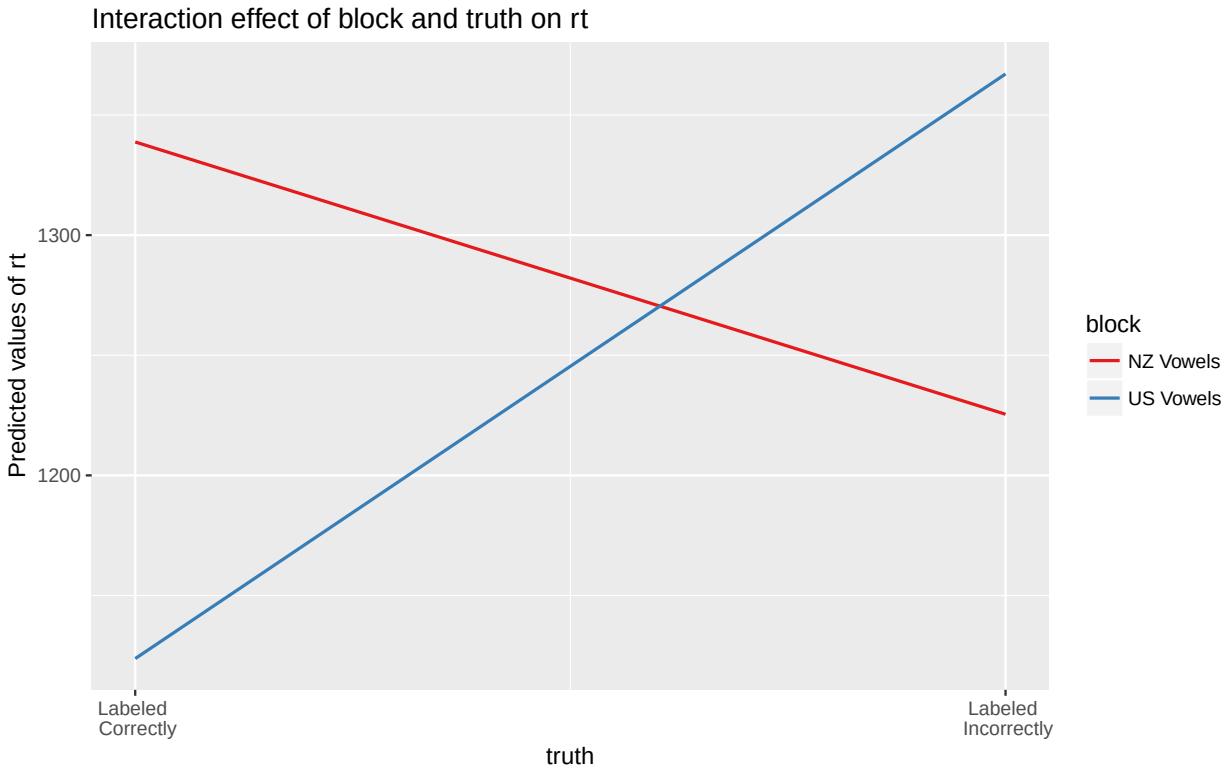

Figure 6.7: Interaction of truth of label on reaction time. Participants answered more quickly when the believed they were listening to their own dialect. This plot shows the coefficients for the interaction of these two terms in a linear mixed effect model which included item as a fixed effect and both response correctness and subject as random effects.

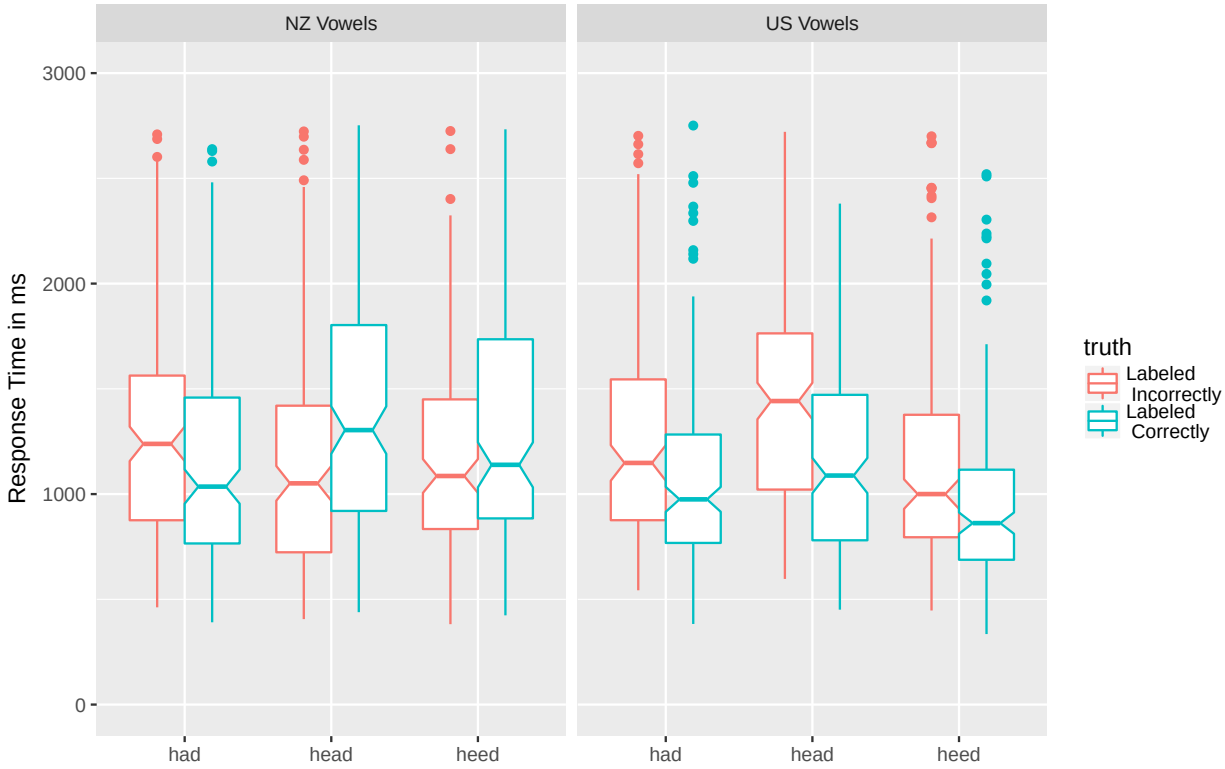

Figure 6.8: Reaction time for tokens, separated by vowel quality, the dialect participants were listening to and whether it was correctly labeled or not. Note that, in general, reaction times are faster when participants believe they are listening to MUSE.

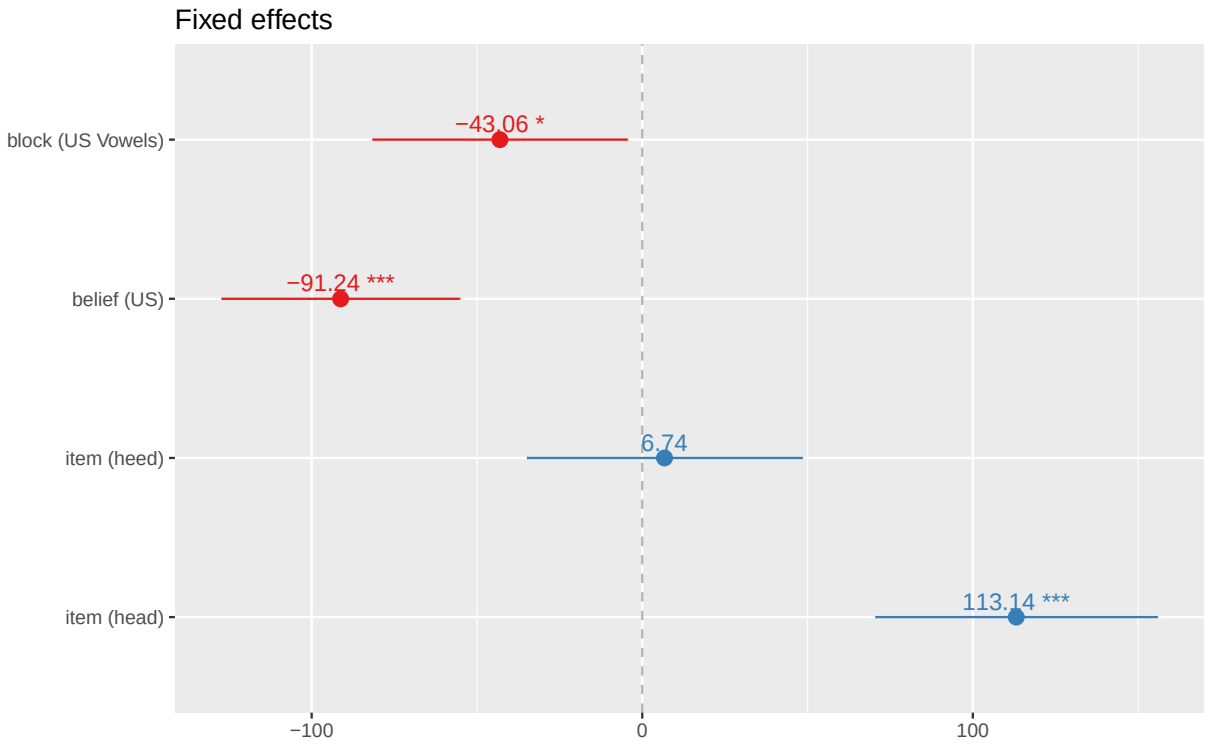

Figure 6.9: Visualization of the fixed effects of item, participant beliefs about which dialect they were listening to and the dialect they were actually listening to on response time. Responses were faster both when the token was from MUSE and when participants believed that it was from MUSE. The effect of belief, however, was greater than that of the actual dialect. The former led to a 95ms reduction in response time, the latter to a 52ms reduction.

The impact of these fixed effects can be seen in Figure 6.9. As discussed previously, there was an effect of item, with "head" being recognized particularly slowly. In addition, both the actual and believed dialects led to changes in response time. For both, participants responded more quickly to tokens from a talker of MUSE. However, belief that a talker was from the US led to quicker responses than a talker actually being from the US. This is further evidence that listeners are using top-down information about a talker's social identify during speech perception.

However, while listener's beliefs did affect the speed with which they responded, it did not affect participant's accuracy. This can be seen in Figure 6.10. This figure shows the relative weights of the same factors as in Figure 6.9. Rather than the output of a linear mixed effect model of response time, however, this model is a binomial Generalized Linear Mixed-Effects model of response correctness, with subject as a random effect. The fact that the error bars for the coefficient of listener belief cross 0 means that there is not sufficient evidence to state that this term is contributing to the model of user accuracy. This is corroborated by the fact that adding listener belief as a fixed effect did not significantly improve model fit ($\chi^2(2, 4029) = 1.38$, $p = 0.24$).

In other words, participants' beliefs change how quickly they respond, but not how often they are correct. Participants were always more accurate when identifying sounds from their own dialect, regardless of their beliefs.

6.3 Discussion

The results presented above provide additional support for the conclusions drawn from Experiment 1. Again, we see evidence that listeners are successfully learning to correctly identify the vowels of NZE, and whether they apply those categories depends on whether they believe they are listening to a talker from that country.

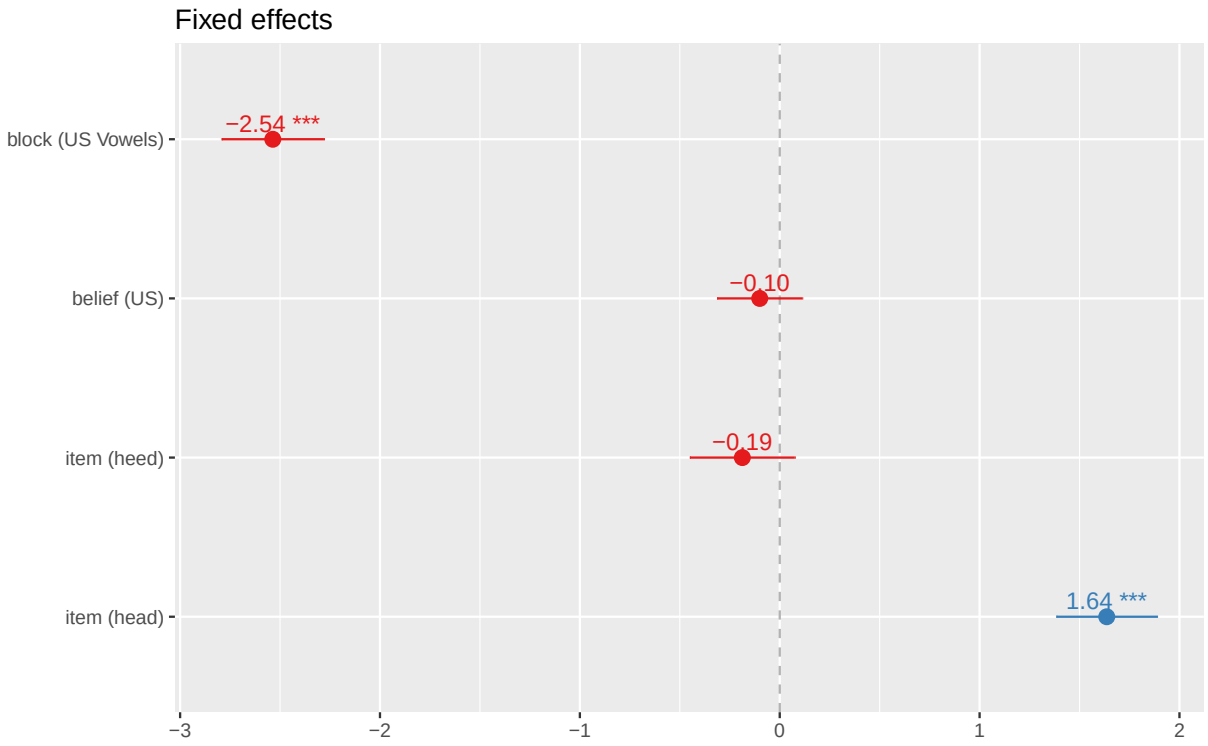

Figure 6.10: Visualization of the fixed effects of item, participant beliefs about which dialect they were listening to and the dialect they were actually listening to on response correctness.

However, with the inclusion of acoustic data from their own dialect, an additional complexity emerges: listeners have a strong bias towards their own dialect. This emerges mainly in participant classifications. Classifications are in line with talkers' own dialect except when they are given data from the training dialect and explicitly told that it is from the training dialect.

Participants' own-dialect bias can be partially explained by the differences in formant dynamics between the two dialects discussed in Section 6.0.1. It is possible that they are sensitive to the formant dynamics of their own dialect, and that this is leading them to (correctly) classify vowels as if they were from their own dialect, even when the social information given to them suggests that they should not. However, since they have had had less exposure to the formant dynamics of NZE, participants are unable to use this subtle acoustic cue to "correct" for the erroneous social information they have been given.

Chapter 7

MODELING LISTENERS' BIAS TOWARDS THEIR OWN DIALECT

This chapter focuses on building an automatic classification model which accurately follows the patterns of human participants' classifications. Based on the experimental data outlined above, any classifier of multi-dialect data should have the following qualities in order to achieve human-like classifications:

1. Classifications should depend on information about the talker's dialect.

2. The classifier should should consider one dialect the "default" and be biased towards it.

7.1 Conditional inference tree models

As in Chapter 5 above, modeling in this section was done using Conditional Inference Trees, as implemented in the partykit R package (Hothorn, Hornik, & Zeileis, 2006).

7.1.1 Data

To maximize parallelism, the same acoustic data was used to train classifiers as the participants were given. For each item, three features were recorded: the talker's regional origin (NZ or US), and the first and second formants as measured at the central point of the steady state of the vowel. Formants are frequency bands of high acoustic energy within the speech signal, and they are a robust cue to the identity of the vowel for human listeners (Wright,

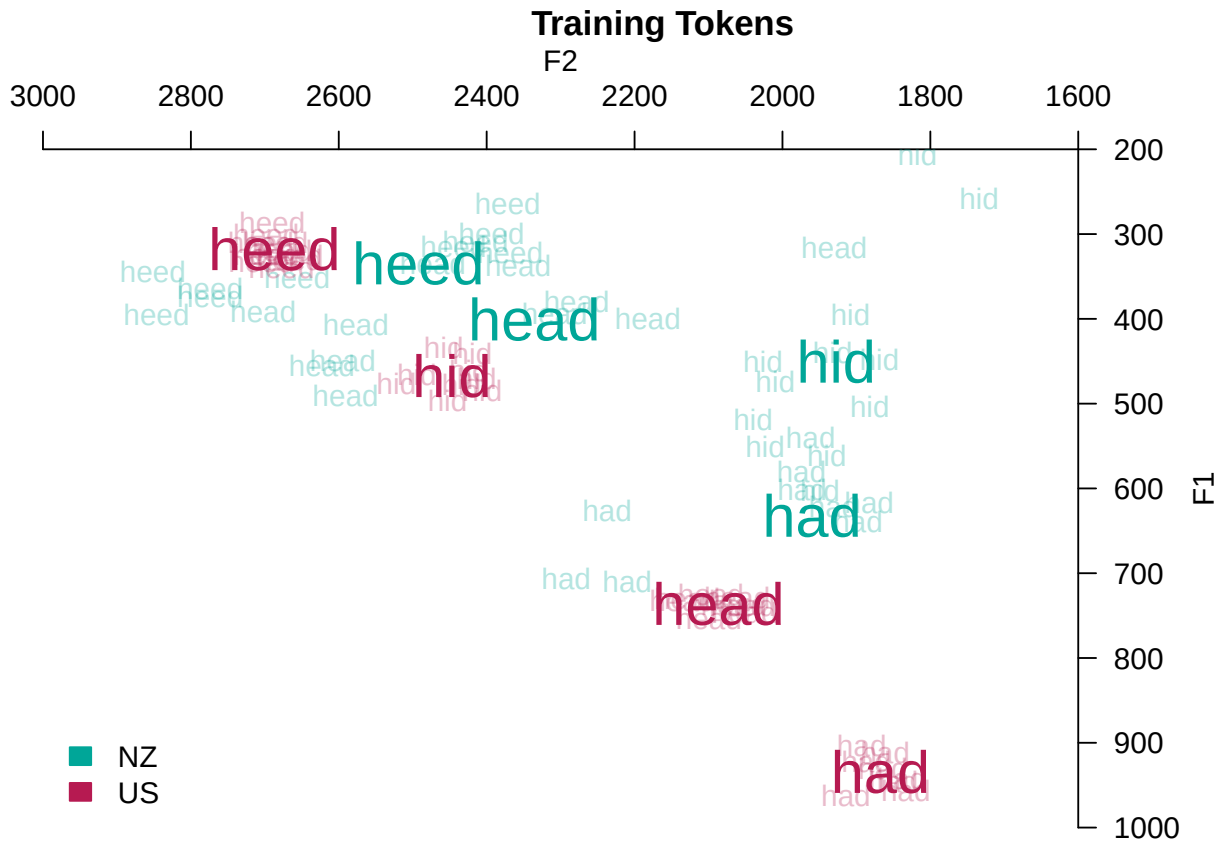

Figure 7.1: Plot of training tokens in a first by second formant space.

2004). The vowels considered in this study are primarily distinguished by the first formant, as can be seen in Figure 7.1.

7.1.2 Bias as exposure

One way to model own-dialect bias is to compare models trained on datasets where there is bias towards one dialect. To simulate this, 4500 conditional inference trees were trained. The proportion of NZE training tokens to MUSE token was varied between 0.02 (only 1 NZE token) to 1.3 (45 NZE tokens), with a uniform distribution. As the proportion of training

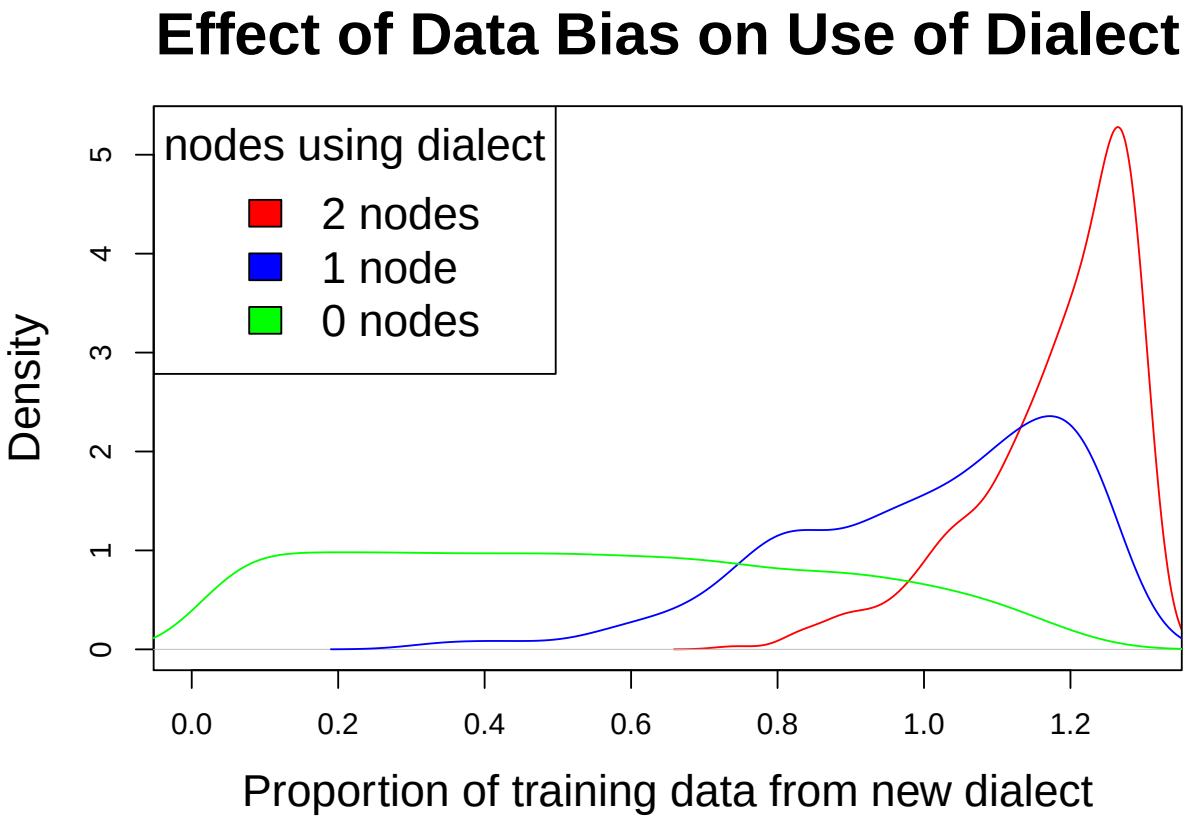

Figure 7.2: Figure showing how bias in training data affects the degree to which dialect information is used during classification. The more data from the second dialect is added to the original dialect, the more nodes in conditional inference tree will split based on dialect.

tokens from the second dialect increased, so did the number of nodes in the conditional inference tree splitting based on talker dialect. This can be seen in Figure 7.2.

A classifier trained on a dataset highly biased towards MUSE tokens was an excellent model of participants' own-dialect bias on mislabeled tokens from their own dialect. This can be seen in Figure 7.3. The classifications on the left are from a model trained on data strongly biased towards MUSE: 90% of the training data for this model came from MUSE. The classifications on the right are from a model trained on a dataset balanced between

MUSE and NZE (50% each). The outputs of both models are based on MUSE data that has been mislabeled as being from NZE. The biased model has classifications much more in line with human participants than the balanced one right. The overall accuracy of biased model (F_1=1.00) was also closer to that of human participants (F_1=0.97) than the accuracy of the unbiased model (F_1=0.85).

However, F_1 only measures the degree to which each classifier agrees with the ground truth. In order to meaningfully compare classifiers with each other, it was necessary to develop a way to compare the agreement of classifiers with each other. In addition, it was necessary to deal with the fact that not all input-output pairs were possible across classifiers; while "hid" tokens were used as inputs for the conditional inference trees, they were never given to human listeners.

For each classifier, the proportion of classifications for each input-output pair such that that pair was possible for both classifiers was calculated. This meant that classifications which were not possible for both classifiers were given a weight of zero. Then, for each input-output pair, the difference was taken between the proportion of the output classification of each classifier. So if neither classifier ever gave a certain output for a given input, the difference would be 0. On the other hand, the more likely one classifier was to assign an output to a given input and the less likely the other was, the greater the difference would be, up to a theoretical difference of .5, if the classifiers never showed the same input-output pairs. The sum of the absolute values of these differences was then taken, the resulting proportion expressed as a percentage. This value represents the percentage of classifications which are not shared between the two classifiers. To aid with interpretation, this value was then subtracted from one, so that it instead represents the percentage of input-output pairs shared between the two classifiers, and it is this which is shown in the following figures. It should be noted that this represents input-output pairs and not necessarily equivalent

classifications of the same tokens.

I also compared balanced and biased models to human participant's classifications of NZE data mislabeled as being MUSE. This can be seen in Figure 7.4. While the balanced model has slightly more overlap with human listeners, it did not capture all the underlying patterns of listener's classifications. In particular, the biased model shows the same confusion pattern as human participants, with "had" mislabeled as "head" and "head" mislabeled as "hid". The model trained on balanced data was more accurate (F_1=0.74, as opposed to F_1=0.48 for the biased model), but, again, pure accuracy isn't necessarily a useful descriptor for how human-like models are: participants were very poor at this task in terms of raw accuracy (F_1=0.56).

But what do these classification models tell us about human learning? Conceptually, we can think of the addition of nodes splitting on dialect as modeling the fact that the more exposure a listener has to a dialect, the more likely they are to make distinctions between that dialect and their own. This is supported by research which shows that listeners are better at identifying dialects the more they have been exposed to them, which has been found in both English (Clopper & Pisoni, 2004; Baker et al., 2009) and Spanish (Díaz-Campos & Navarro-Galisteo, 2009).

While greater exposure to multiple dialects may improve comprehension accuracy overall, it does come at a cost–at least for human listeners. It has been previously found that lexical selection is slower for listeners who have been exposed to multiple dialects (Clopper & Walker, 2016). While all participants in this experiment were exposed to multiple dialects, we can compare their response times for "heed" (which was very similar across both dialects) to those for "had" and "head" tokens (which diverged). As discussed in Section 6.2.1.1 above, response times for human participants were significantly faster for "heed" tokens than "head" or "had" tokens. A one-way ANOVA of log-transformed response times in milliseconds

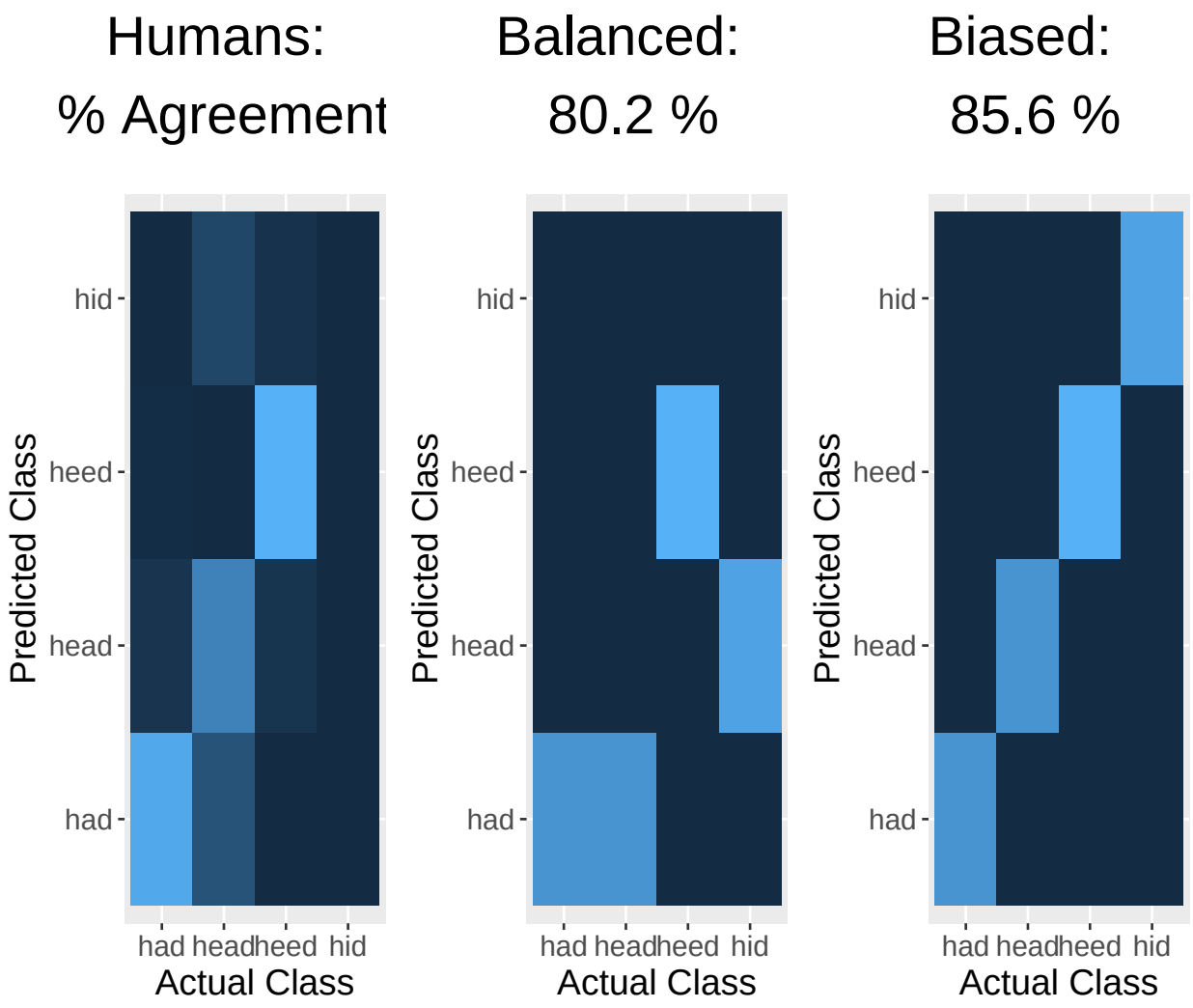

Figure 7.3: Classifcations of human participants (on left) compared to classifications by conditional inference trees trained on data balanced between MUSE and NZE (in center) or biased towards MUSE (on right). All three were given MUSE tokens mislabeled as NZE. In this case, the biased classifier is a better behavioral model: 85.6% of the classifications from the biased model overlapped with human classifications.

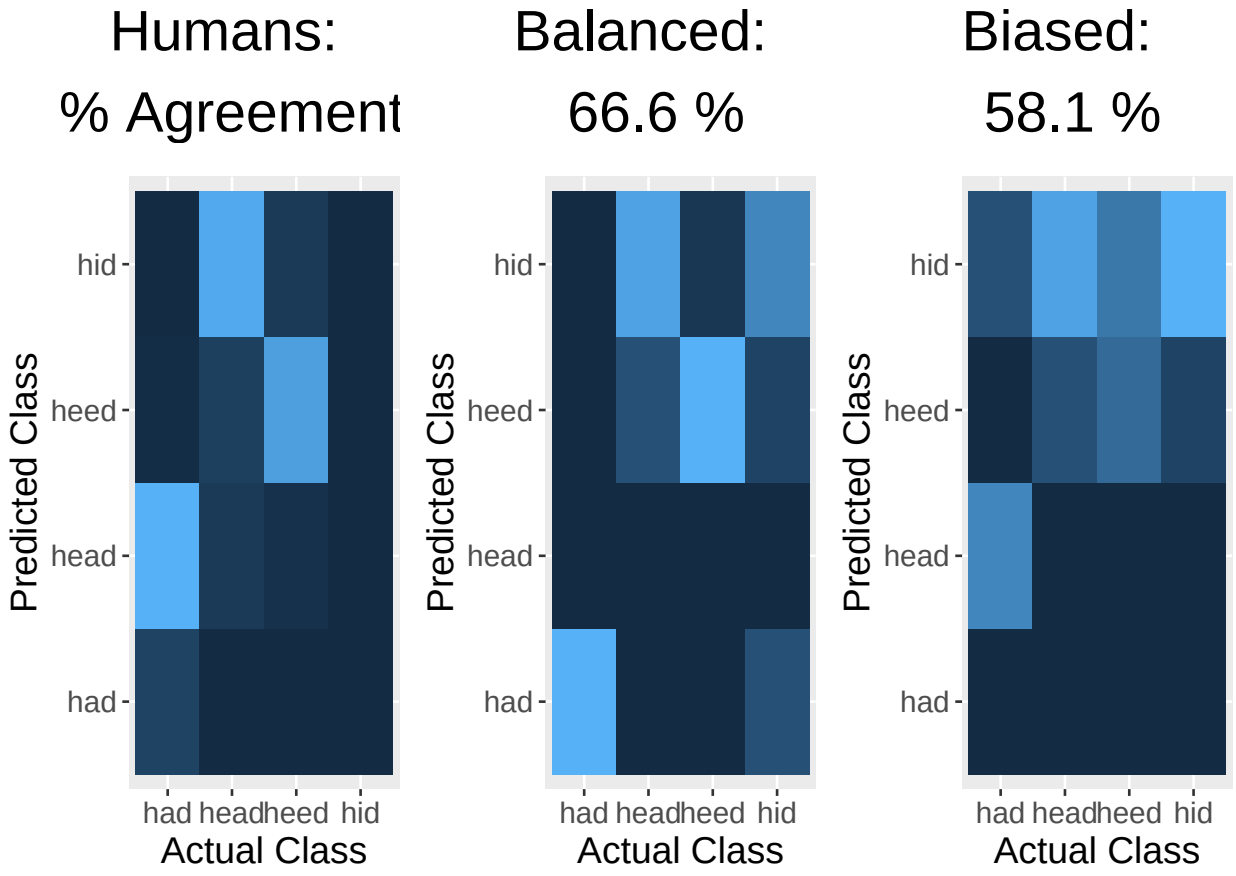

Figure 7.4: Classifications of NZE tokens mislabeled as MUSE by classifiers trained on data biased towards MUSE (left) or balanced between MUSE and NZE (right). T balanced training data resulted in classifications that were more accurate and had more overlap with the human listener's classifications, but the biased model captured the human listener's mis-classification of "had" as "head".

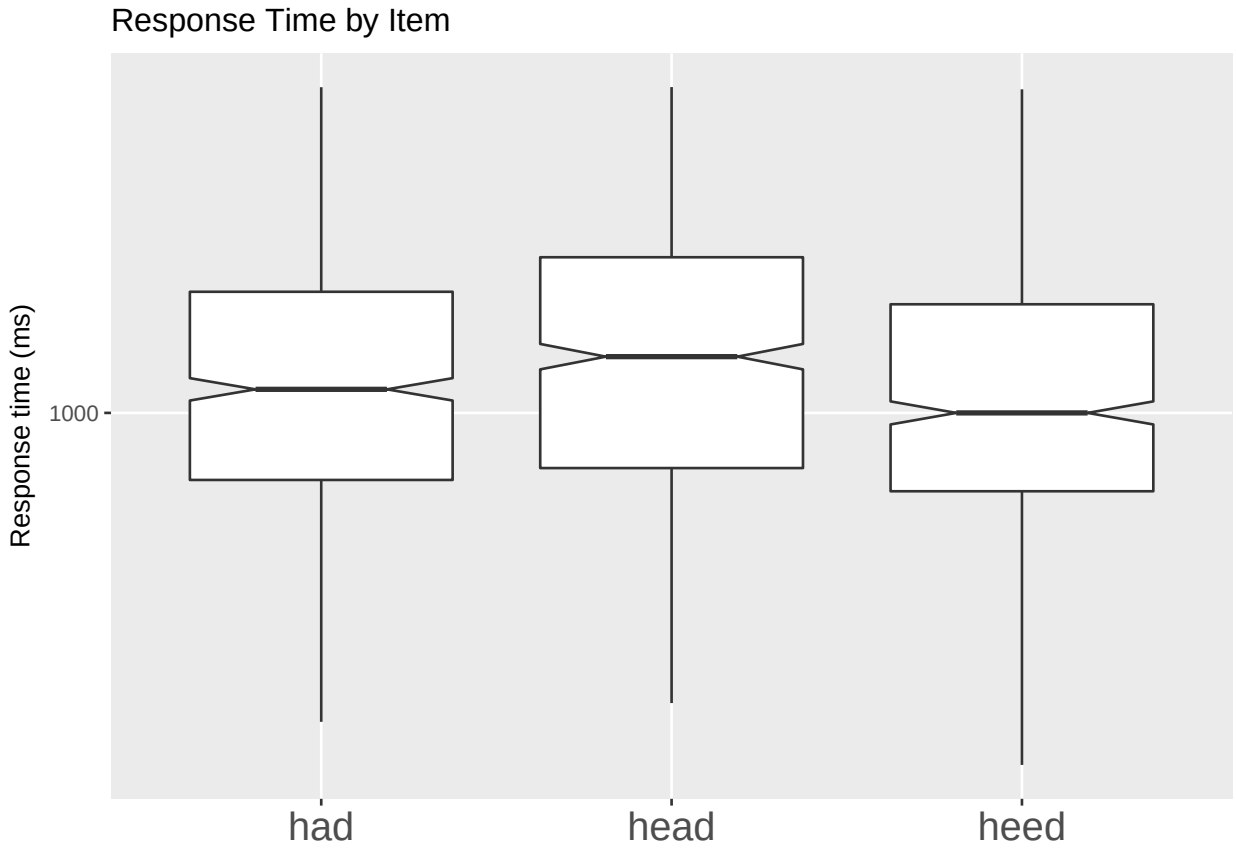

Figure 7.5: Log response time, in milliseconds, by token type. Note that “heed" tokens, which were similar across dialects and thus didn’t require dialect-specific learning, were categorized most quickly.

found a significant effect of word ($F(2,4029)=20.63$, $p > 0.01$)). Response times for “heed" ($\mu = 1168.38, \sigma = 541.13$) were the fastest, followed by “had" ($\mu = 1196.62, \sigma = 510.59$) and “head" ($\mu = 1298.16, \sigma = 563.39$). So response times for tokens confusable across dialects were slower than for tokens that were similar across dialects. This can be seen in Figure 7.5.

The models presented here help to quantity the fact that listeners’ have a strong bias towards their own dialect when hearing it. As discussed in Chapter 6 above, while listeners may not be explicitly identifying that the speech they’re hearing is their own dialect, differ-

ences in formant dynamics between the dialects may be robust enough cues to vowel quality to account for this bias. When listening to a dialect that they had only recently encountered, however, listeners' gave more equal weights to both their own dialect and the dialect they were learning.

7.1.3 Differences between this and other learning models

This chapter modeled the experimental finding that human listeners maintain a perceptual bias towards their own dialect even after successful perceptual learning. These findings are not surprising given previous findings of social bias towards ones own dialect (Coupland & Bishop, 2007; Floccia et al., 2006), as well as findings in the speech perception literature.

Early work on the perceptual magnet effect (Kuhl, 1991), for example, showed that listeners have a perceptual preference for vowels they are more familiar with, which was attributed to the location of prototypes. Later work on episodic learning casts this preference in terms of a higher density of exemplars in this area (Johnson, 1997; Pierrehumbert, 2001).

Though the modeling results discussed here do fit well with the larger claims of exemplar theory, they are not an explicit extension of it. Note that the balanced conditional inference tree model is not explicitly conditioned on the biased models; each is a separate, independent model. This differs from the Bayesian belief-updating modeling often employed in exemplar theory. Learning can be modeled in this framework in much the same way, however. Training a new model on a data set that is a superset of the previous training set would allow for "updating" of the model.

The model presented here was weighted to show own-dialect bias. However, there is nothing to stop the weighting from going the other way, which would suggest that it is possible for a listener to be biased away from their own dialect. Given that this has previously been observed in phonetics research, this is a feature rather than a bug. With enough exposure

to a second dialect, especially if their attitude towards the second dialect is positive, talker's production and perception can both shift towards the new dialect (Evans & Iverson, 2007; Sumner & Samuel, 2009).

The effect of bias in training data on the classification of conditional inference tree models closely parallels human listeners' bias towards their own dialect. While this is desirable from a behavioral modeling perspective, dialectically-biased training data also resulted in lower accuracy on the dialect it was biased against. This suggests that systems trained primarily on one dialect should make the same types of errors on other dialects as a talker of the training dialect would. In this case, an ASR system trained on MUSE data would make systematic errors on NZE.

7.2 Implications for automatic speech recognition

The model discussed above provides an alternate way to include talker dialect during automatic speech recognition. Rather than training separate models on each dialect (Telaar & Fuhs, 2013; Najafian, Safavi, Hanani, & Russell, 2014), dialect can be included within decision-tree based acoustic models (S. J. Young, Odell, & Woodland, 1994), so that the classification of each phone considers talker dialect as well as its acoustic properties. This may be especially helpful for closely related dialects or varieties that share phones.

A second advantage of dialect tags during phone classification is that it improves accuracy beyond the gain made by simply varying training data. Even more crucially, it improves performance on the target dialect without reducing the accuracy of the base dialect, which has been a problem for other accent adaptation systems (Elfeky, Moreno, & Soto, 2015; Nallasamy, 2016).

7.2.1 Adding dialect as as feature improves performance beyond balancing training data

In order to evaluate the effect of specifically adding demographic labels as opposed to just adding additional data, three separate conditional inference tree models were trained. Each was cross-validated with the same testing dataset. The training datasets, however, varied. The first model was trained only on US English. The second was trained on a training set that included the US English used to train the first model and an equal amount of NZ English data. For this model, however, the training data was not labeled with its dialect of origin. The third model was trained on the same data as the second, but with the correct dialect tags included. For this final classifier, correct dialect labels were used during training and also given as a feature at the time of classification; automatically identifying which dialect a talker is using is beyond the scope of this project.

The accuracy for these classifiers on both dialects can be seen in Table 7.1. When the classifier is trained on only US English, the accuracy on NZ English is low (53%), but still above chance (25%, since this is four-way classification task). Adding a second dialect to the training data does reduce accuracy for the base dialect (albeit only very slightly), but also results in a 17% improvement in accuracy for the target dialect. Adding dialect labels results in another 6% gain in accuracy, while at the same time remaining highly accurate for the base dialect. There is no reason to suppose that correct dialect labels would reduce accuracy for either dialect, but this does provide validation for the benefits of adding dialect labels.

7.2.2 Possible draw-backs

This increased accuracy may, however, come with a cost. The largest hurdle is that the model proposed here assumes that talker's dialect is accurately labeled. Doing so requires accurate dialect identification, which is an active area of research, but outside the scope of

Model	Accuracy, US talkers	Accuracy, NZ talkers
Baseline, trained on only US data	**0.91**	0.53
Trained on both US & NZ data, no demographics	0.9	0.7
Trained on both, with demographic data as a feature	0.9	**0.76**

Table 7.1: Table showing accuracy of conditional inference tree classification of phones, with the highest accuracy on each dialect bolded.

this dissertation **(Biadsy, 2011, for review)**.

The second drawback of this proposal is that it requires all training data be accurately labeled for dialect of origin. If human-provided labels are available that would be ideal, however that is impractical for large datasets. Again, this would require accurate automatic dialect identification and labeling if a large-scale system is to be implemented.

Chapter 8
CONCLUSION

This dissertation has drawn together work on system evaluations, experimental investigations of human learning and machine learning models of human learning to answer one central question: how can we help automatic speech recognition systems cope with the sociolinguistic variation inherent in all language? Evaluation of current commercial ASR systems has shown that this is still a hurdle for current systems, and that as it stands the users which suffer the brunt of the negative effects are ones who already face other disadvantages.

The experiments outlined here suggest one strategy that human listeners use when encountering new dialects: relying on explicitly extra-linguistic information about what dialect a speaker is likely to use. However, this is only true when they are listening to a dialect they have limited experience with. When listening to their own dialect, listeners disregard extra-linguistic information that would lead them to make less accurate classifications.

The bias that listeners had towards their own dialect was accurately modeled by adjusting the volume of training data given to a classifier from each of two dialects. Listeners' classifications for their own dialect are better modeled by a classifier trained on data mostly from their dialect. On the other hand, listeners' classifications for dialect they have recently encountered are better modeled with a classifier trained on a balanced mix of speech both dialects. More than just adding data from each dialect, labeling which dialect each speech sample in the training data came from and then providing dialect labels as features at the time of classification greatly improved the accuracy of the classifier on the new dialect.

Taken together, the experimental and modeling results suggest one thing: social data

about which dialect a speaker is likely to be using is more useful when listeners have less experience with a dialect. For human listeners, this makes sense from a perceptual learning perspective. Once they have a large body of experience with a dialect, listeners are more likely to engage in differentiation using fine-grain cues that might escape the notice of a listener less familiar with that language variety.

This insight also has implications for automatic speech recognition. In particular, it suggests that using training and test data labeled for the user's likely dialect might result in larger gains for low-resource varieties of a language, especially if data from other varieties of that language is already available.

8.1 A continuing need for accent adaptation

The evidence laid out in Chapter 3 shows the continuing need for work on accent and dialect adaptation in automatic speech recognition systems. While the differences in accuracy observed between male and female talkers was only observed in the first study, and might be attributable to variation in signal to noise ratio, the differences between talkers from different dialects was replicated across systems. While non-native accents vary in sometimes unpredictable ways, native talkers' dialects vary in stable ways that automated systems should be able to correctly adapt to.

One might ask, if these differences in accuracy arise from unbalanced training data, are lower accuracy rates acceptable for talkers from minority groups, given that there fewer of them in the overall population? I would argue that they are not, especially given the evidence I have provided above that it is possible to increase accuracy in phone identification for the target group without reducing it for the main group.

In addition, the fact that some talkers enjoy more accuracy when using automatic speech recognition (ASR) systems is especially troubling given that the groups who are recognized

most accurately (white talkers of standardized American English) are those whose social position is most privileged. This is a problem because it makes technologies which incorporate ASR less accessible to groups already facing marginalization. As these technologies are increasingly incorporated into professional contexts, in particular hiring (Shahani, 2015; L. Morrison, 2017) and the use of ASR tools in jobs requiring dictation, especially in medical contexts (Ali et al., 2007), this may further exacerbate existing social inequalities. In hiring contexts in particular, these differences in algorithmic performance between social groups may represent unintentional but still harmful discrimination (Ajunwa, Friedler, Scheidegger, & Venkatasubramanian, 2016).

8.2 The use of social information in applying perceptual learning

To some extent, this work replicates earlier findings of the importance of social beliefs in perception by showing that listener's social beliefs can categorically change their perception. However, this effect was only found for the dialect that listeners had less experience with. This suggests that the amount of previous experience a listener has had with a dialect plays an important role in the degree to which they use extra-linguistic information during perception.

Which factors affect the likelihood that a listener will use extra linguistic social information during perception? Both the amount of exposure a listener has had to a dialect and the amount of acoustic information available seem to affect the degree to which giving a listener explicit social information about a talker will affect their speech perception.

The amount of exposure a listener has had to a dialect. The more experience a listener has had with a dialect, the more likely they are not to need to rely on extra linguistic information. Listeners who have had extensive experience with a dialect, or "fluent listeners" (Sumner & Samuel, 2009), are likely to extract all information they require for

speech perception from the signal. When listeners have had less experience with a variety, and are thus more uncertain about the sounds they are identifying, they are more likely to make use of extra-linguistic information.

The amount of acoustic information available. The more acoustic information a listener is given, the less they will need to rely on extra-linguistic acoustic cues for perception. This will also help to explain a number of studies which have failed to find an effect of social priming. Squires, for instance, failed to find the expected social priming for the stigmatized NP + "don't" structure (Squires, 2013). This may be partially due to the fact that the audio stimuli used in this case were entire sentences. Similarly, Lawrence failed to find social priming for the stereotyped BATH and STRUT vowels in a study where the target words were played at the end of a recorded sentence (Lawrence, 2015). Juskan replicated this failure to find social priming in his dissertation, where, once again, the target tokens were presented in full sentences (Juskan, 2016). These results are unsurprising if, as I posit, listeners familiar with a variety are capable of extracting relevant social information from from very little acoustic information. If 150ms worth of a single vowel contains enough information for a listener to disregard information meant to facilitate social priming, then entire sentences certainly do as well.

8.2.1 The use of social information in day-to-day interaction

As outlined above, there are a limited number of situations where different social information about the individual who made the same set of speech sounds will categorically change speech perception. So how does this affect speech communication in day-to-day interactions in the "real world"? This is a hard question, and one which the work in the dissertation is insufficient to answer. However, there are some specific interactional situations where social information

alone may be enough to change a listener's precepts.

One situation is when first meeting a new talker. In this situation, a listener will not have previous experience to draw upon and will not be able to make use of talker-specific learning. Listeners in this situation may rely more heavily on social information, especially if that information is that the person may belong to a group which uses a language variety different from the listener's own and whose language the listener has previously encountered.

In addition, it seems likely that social information will be weighted more highly when there is less acoustic information available. I would predict, for example, that listeners will show a greater reliance on information about the social identity of the speakers if the formant structure is masked or degraded in some way. This is in keeping with research that suggests that listeners rely more on context clues when the speech signal is degraded (Miller, Heise, & Lichten, 1951; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005).

Finally, if we broaden "social information" to mean not just information on the speaker themselves, but also their social environment, this information may be more highly weighted for talkers who are bi- or multidialectal. Bidialectal speakers have been found to fluidly switch between dialects depending on the larger social context of their interactions (Blom & Gumperz, 2000; Davies, 2007), and knowledge of this larger social context and which dialect features a talker is likely to use in it could have measurable effects on speech perception.

8.2.2 Own-dialect advantage in learning novel dialects

Given earlier work, it is not surprise to find that a talker's native dialect affects their perceptions. While one early study failed to find robust differences in perception between talkers from Chicago and Oklahoma (Fox, 1974), many subsequent studies have found robust effects of a listener's own dialect on their speech perception. Zanten and van Heuven, for instance, found that listeners who spoke different dialects of Indonesian showed differences in both la-

beling and acceptability judgments of vowel tokens (Zanten & Heuven, 1984). For American English, multiple studies have found that listeners from different dialect backgrounds show shifts in the perceptual boundaries between vowels (Kendall & Fridland, 2012; Fridland & Kendall, 2012). Other work has found that not only a listener's own dialect, but also their geographic mobility, affects their perceptions (Bowie, 2000; Evans & Iverson, 2007; Sumner & Samuel, 2009). There is also evidence that listener's dialect background affects what types of lexical confusions they're likely to make (Clopper, Pierrehumbert, & Tamati, 2010).

There is also electrophysiological evidence for differences in perception based on the listener's dialect background. (Brunellière, Dufour, Nguyen, & Frauenfelder, 2009) found that talkers of a dialect of French that has a distinction between /e/ and /ɛ/ but none between /ø/ and /y/ produced a mismatch negativity only when presented with /e/ and /ɛ/, though an acoustically comparable difference existed between /ø/ and /y/. This shows that even on a neural level, a listener's dialect affects their perception of speech.

8.2.3 Identifying what dialect a talker is using

The work presented here does skirt of the central problems of both accent adaptation for ASR and perceptual dialect learning: how do you identify what dialect a talker is using?

There is a large body of work on listeners' ability to correctly identify a talker's dialect based on the acoustic information they are presented with. Naive listeners can use acoustic cues to distinguish broad dialect classifications (Clopper, 2000). They also rate dialects geographically closer to their own as being more like theirs, but this effect is weaker for geographically mobile listeners (Clopper, 2004). There is also some evidence that listeners are better at identifying dialects from within their own country (Ikeno & Hansen, 2006). However, even when talkers cannot correctly identify the dialect they're listening to, they can still understand the speech they're hearing (Clopper & Bradlow, 2008).

So listeners can simultaneously struggle to correctly identify what dialect a talker is using while still easily understanding their speech. This suggests that explicit dialect labeling is not necessary for speech perception. This is in line with the results presented above, which show that listeners who are experienced with a dialect will correctly reject incorrect dialect labels and accurately categorize sounds based on very sparse acoustic information. Rather than identifying and categorizing the relevant social characteristics of a talker, listeners may instead be relying on their previous experience of possible variants. While top-down information (in this case dialect labels) can play a role in speech perception, they are most likely to do so when listeners have less experience with the dialect they are hearing.

8.3 Implications and future work

In addition to expanding our understanding of how listeners use both linguistic and extra-linguistic information when learning novel dialect features, this dissertation has direct implications for making ASR accent adaptation more human-like, as well as suggesting a number of possible future directions for additional research.

8.3.1 Automatic speech recognition

What are the implications of the behavioral results and modeling presented here for making automatic speech recognition more human like?

The first is that social inference about talkers is necessary for human-like perception, but only under certain conditions. In particular, it can be useful when there is limited acoustic information about the dialect being adapted towards and when there is extra-linguistic information about a talker's probable language variety, such as GPS information (Chelba et al., 2015; Elfeky et al., 2015; Ye et al., 2016).

The second is that, especially for language varieties for which a large amount of acoustic

training data is available, explicitly labeling a specific talker's dialect may not be be necessary for human-like dialect adaptation. In fact, it may be detrimental. An individual talker has access to a range of possible variants, and may switch between them based on, for instance, their audience (Bell, 1984) or a desire to shift their production towards or away from their interlocutor (Giles et al., 1991). In addition, an individual talker may use more than one dialect, and code-switch between them (Milroy & Muysken, 1995, for discussion). Thus statically assigning a talker one dialect label may not accurately reflect their use of language. However, it should be noted that there is no reason why dialect should be labeled at the level of the speaker. If it is instead labeled at the level of the phone or word, then the output could be generated by comparing hypotheses across dialect models/dialect labels, in an approach similar to that taken by Soto and co-authors (Soto, Siohan, Elfeky, & Moreno, 2016).

8.4 Directions for future work

One clear next step in this research program is to work on an implementation of accent adaptation for ASR which incorporates these findings. In addition to implementing such a system, there a number of possible future directions in terms of cognitive and behavioral modeling.

On a very broad level, these question relate to the interplay of top-down (extra linguistic social knowledge) and bottom-up (acoustic) information during speech perception. There is ample evidence that listeners are using top-down information about a talker's social identity during perception, it is clearly still also the case that listeners can use bottom-up acoustic information to arrive hypothesize about the talker's social identity. When is each of these types of information used, and what is the balance between them?

How much acoustic information do people need to reliably identify a dialect, or more generally that a talker is not using a dialect they are familiar with? This question is of

particular interest in light of the failure to find social priming effects when listeners are presented with large amounts of acoustic information, such as entire sentences. Exactly how much acoustic information can a listener hear before developing certainty about a the talker's social identity? Does degrading the acoustic information, for example by embedding it in noise, increase the amount which listeners need to be exposed to before they make social inference?

On a related note, are some parts of the speech signal more informative than others? For instance, are differences in vowel space more salient than consonant differences, such as flapping? Is prosody perceptually important for dialect identification? When these cues suggest contrasting information, which do listeners rely on more? It is likely that which differences are most salient for any two pairs of dialects will depend on the dialects themselves, but are there general classes of sounds which are more informative than others?

In addition, the work presented here has looked at social beliefs at one remove (as they are expressed in the identification of speech sounds) rather than explicitly asking for social evaluations or for listeners to label a talker's dialect. Doing so may reveal additional useful insights.

8.5 Conclusion

This dissertation has made three main contributions. The first is to document remaining sociolinguistic biases in state-of-the-art commercial automatic speech recognition systems. The second is to shed new light on how listeners use linguistic and extra-linguistic information during speech perception; social information is more likely to directly affect perception when listeners are listening to a variety they have less exposure to. This may be the result of listeners' sensitivity to subtle acoustic cues, such as formant dynamics, which are present in even very short snippets of speech. Finally, this dissertation has modeled the asymmetry in

how by extra-linguistic information is used when listnening to different dialects by training conditional inference trees on training datasets which contain more data from one dialect than another. The modeling results suggest that using extra-linguistic information could be most useful for automatic speech recognition systems when less acoustic data is available for one variety than another.

References

Abdulla, W., Kasabov, N., & Zealand, D.-N. (2001). Improving speech recognition performance through gender separation. *Changes*, *9*, 10.

Adank, P., Nuttall, H. E., Banks, B., & Kennedy-Higgins, D. (2015). Neural bases of accented speech perception. *Frontiers in Human Neuroscience*, *9*(558).

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*(5), 3099–3107.

Ajunwa, I., Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2016). Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*.

Ali, S., Siddiqui, K., Safdar, N., Juluru, K., Kim, W., & Siegel, E. (2007). Affect of gender on speech recognition accuracy. In *American journal of roentgenology* (Vol. 188).

Aubanel, V., & Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*, *52*(6), 577–586.

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823–2833.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180.

Bailey, G., & Thomas, E. (1998). Some aspects of African-American vernacular English phonology. *African-American English: structure, history, and use*, 85.

Baker, W., Eddington, D., & Nay, L. (2009). Dialect identification: The effects of region of origin and amount of experience. *American Speech*, *84*(1), 48–71.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(02), 145–204.

Bent, T., Bradlow, A. R., & Wright, B. A. (2006). The influence of linguistic experience on the

cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 97.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786.

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification a McGurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener?s native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775-794.

Biadsy, F. (2011). *Automatic dialect and accent recognition and its application to speech recognition*. Unpublished doctoral dissertation, Columbia University.

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868*.

Blom, J.-P., & Gumperz, J. J. (2000). Social meaning in linguistic structure: Code-switching in Norway. *The Bilingualism Reader*, 111–136.

Bock, B., & Shamir, L. (2015). Assessing the efficacy of benchmarks for automatic speech accent recognition. In *Proceedings of the 8th international conference on mobile multimedia communications* (pp. 133–136).

Boersma, P., et al. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bogue, D. J., Anderton, D. L., & Barrett, R. E. (2010). *The population of the United States*. Simon and Schuster.

Bowie, D. (2000). *The effect of geographic mobility on the retention of a local dialect*. Unpublished doctoral dissertation, University of Pennsylvania.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Iv. some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Brunellière, A., Dufour, S., Nguyen, N., & Frauenfelder, U. H. (2009). Behavioral and electrophysiological evidence for the impact of regional variation on phoneme perception. *Cognition*, *111*(3), 390–396.

Chambers, J. K. (1995). *Sociolinguistic theory*. Blackwell.

Champely, S. (2016). pwr: Basic functions for power analysis [Computer software manual]. Available from `https://CRAN.R-project.org/package=pwr` (R package version 1.2-0)

Chelba, C., Zhang, X., & Hall, K. B. (2015). Geo-location for voice search language modeling. In *Interspeech* (pp. 1438–1442).

Chen, K.-t., & Wang, H.-m. (2001). Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation. In *IEEE international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 317–320).

Clopper, C. G. (2000). Some acoustic cues for categorizing American English regional dialects: An initial report on dialect variation in production and perception. *Research on Spoken Language Processing. Progress Report o. 24 (2000)*, 43–65.

Clopper, C. G. (2004). *Linguistic experience and the perceptual classification of dialect variation*. Unpublished doctoral dissertation, Indiana University.

Clopper, C. G. (2007). Effects of dialect variation on speeded word classification. *The Journal of the Acoustical Society of America*, *121*(5), 3189–3190.

Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and speech*, *51*(3), 175–198.

Clopper, C. G., Pierrehumbert, J. B., & Tamati, T. N. (2010). Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise. *Laboratory Phonology*, *1*(1), 65–92.

Clopper, C. G., & Pisoni, D. B. (2004). Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change*, *16*(01), 31–48.

Clopper, C. G., & Walker, A. (2016). Effects of lexical competition and dialect exposure on phonological priming. *Language and Speech*, 0023830916643737.

Cohen, J., et al. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cole, R. A., Cooper, W. E., Singer, J., & Allard, F. (1975). Selective adaptation of English consonants using real speech. *Perception & Psychophysics*, *18*(3), 227–244.

Cole, R. A., Noel, M., Lander, T., & Durham, T. (1995). New telephone speech corpora at CSLU. In *Eurospeech.*

Coupland, N., & Bishop, H. (2007). Ideologised values for British accents. *Journal of Sociolinguistics*, *11*(1), 74–93.

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204.

Current population survey, annual social and economic supplement (Tech. Rep.). (2011). United States Census Bureau. Available from `https://www.census.gov/cps/data/`

Cutler, C. A. (1999). Yorkville crossing: White teens, hip hop and African American English. *Journal of Sociolinguistics*, *3*(4), 428–442.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech

perception: Adjusting the signal or the representations? *Cognition*, *108*(3), 710–718.

Davies, C. E. (2007). Language and identity in discourse in the American South: Sociolinguistic repertoire as expressive resource in the presentation of self. *Selves and Identities in Narrative and Discourse. Amsterdam, The Netherlands: John Benjamins*, 71–88.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222.

Díaz-Campos, M., & Navarro-Galisteo, I. (2009). Perceptual categorization of dialect variation in Spanish. In *Selected proceedings of the 11th hispanic linguistics symposium* (pp. 179–195).

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass*, *4*(7), 473–480.

Droua-Hamdani, G., Selouani, S.-A., & Boudraa, M. (2012). Speaker-independent ASR for Modern Standard Arabic: effect of regional accents. *International Journal of Speech Technology*, *15*(4), 487–493.

Eckert, P. (1989a). *Jocks and burnouts: Social categories and identity in the high school.* Teachers College Press.

Eckert, P. (1989b). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, *1*(03), 245–267.

Eckert, P. (2008). Where do ethnolects stop? *International Journal of Bilingualism*, *12*(1-2), 25–42.

Eisenstein, J. (2017). Written dialect variation in online social media. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *Handbook of dialectology.* Wiley.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech

processing. *Perception & Psychophysics*, *67*(2), 224–238.

Elfeky, M., Moreno, P., & Soto, V. (2015). Multi-dialectical languages effect on speech recognition: Too much choice can hurt. In *International conference on natural language and speech processing.*

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America*, *115*(1), 352–361.

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, *121*(6), 3814–3826.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley & Sons.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1276.

Fought, C. (2002). *Chicano English in context.* Springer.

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, *1*(1), 5–39.

Fox, R. A. (1974). An experiment in cross-dialect vowel perception. In *Tenth regional meeting–Chicago Linguistic Society* (pp. 178–185).

Freeman, V., Chan, J., Levow, G.-A., Wright, R., Ostendorf, M., Zayats, V., et al. (2014). ATAROS technical report 1: Corpus collection and initial task validation. *U. Washington Linguistic Phonetics Lab.*

Fridland, V. (2003). 'Tie, tied and tight': The expansion of /ai/ monophthongization in African-American and European-American speech in Memphis, Tennessee. *Journal of*

Sociolinguistics, *7*(3), 279–298.

Fridland, V., & Kendall, T. (2012). Exploring the relationship between production and perception in the mid front vowels of US English. *Lingua*, *122*(7), 779–793.

Ganitkevitch, J. (2005). Speaker adaptation using maximum likelihood linear regression. In *Rheinisch-Westfalische Technische Hochschule Aachen Automatic Speech Recognition Course.*

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., et al. (1993). TIMIT acoustic-phonetic continuous speech corpus. *Linguistic data consortium, Philadelphia*, *33*.

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE transactions on speech and audio processing*, *2*(2), 291–298.

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice*, *19*(4), 544–554.

Georgakis, C., Petridis, S., & Pantic, M. (2016). Discrimination between native and non-native speech using visual features only. *IEEE transactions on cybernetics*, *46*(12), 2758–2771.

Gibson, E. J. (1969). Principles of perceptual learning and development.

Gibson, E. J., & Walk, R. D. (1960). *The "visual cliff"* (Vol. 1). WH Freeman Company.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585–612.

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates.

Speech Communication, *52*(3), 181–200.

Gordon, M. J. (2001). *Small-town values and big-city vowels: A study of the Northern Cities Shift in Michigan*. Duke Univ Press.

Grieve, J. (2016). *Regional variation in written American English*. Cambridge University Press.

Harnsberger, J. D. (2001). The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis. *Journal of Phonetics*, *29*(3), 303–327.

Harrenstien, K. (2009). Automatic captions in YouTube. *The Official Google Blog*, *11*.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892.

Hay, J., Drager, K., & Warren, P. (2010). Short-term exposure to one dialect affects processing of another. *Language and speech*, *53*(4), 447–471.

Hay, J., Maclagan, M., & Gordon, E. (2008). *New Zealand English*. Edinburgh University Press.

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

Hirschberg, J., Litman, D., & Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, *43*(1), 155–175.

Honorof, D., McCullough, J., & Somerville, B. (2000). Comma gets a cure: A diagnostic passage for accent study. *Retrieved March*, *20*, 2017.

Hothorn, T., Hornik, K., & Zeileis, A. (n.d.). party: A laboratory for recursive part (y) itioning. R package version 0.9-9999. 2011. *URL: http://cran. r-project. org/package= party (1 December 2010)*.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional

inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Hothorn, T., & Zeileis, A. (2014). *partykit: A modular toolkit for recursive partytioning in R* (Tech. Rep.). Working Papers in Economics and Statistics.

Houtgast, T. (1995). Psychophysics of speech and speech-like stimuli. *The Journal of the Acoustical Society of America*, *97*(5), 3259–3259.

Huang, R., & Hansen, J. H. (2007). Unsupervised discriminative training with application to dialect classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, *15*(8), 2444–2453.

Huffman, S. (2014, October). OMG! mobile voice survey reveals teens love to talk. *Google Official Blog*. Available from `https://googleblog.blogspot.co.uk/2014/10/omg-mobile-voice-survey-reveals-teens.html`

Humphries, J., Woodland, P. C., & Pearce, D. (1996). Using accent-specific pronunciation modelling for robust speech recognition. In *Fourth international conference on spoken language* (Vol. 4, pp. 2324–2327).

Ikeno, A., & Hansen, J. H. (2006). Perceptual recognition cues in native English accent variation: "listener accent, perceived accent, and comprehension". In *IEEE international conference on acoustics, speech and signal processing* (Vol. 1, pp. I–I).

International toursim snapshot (Tech. Rep.). (2015, June). Tourism Australia. Available from `http://www.tourism.australia.com/documents/Statistics/TACP9963\_International\_Tourism\_Snapshot\_2015\_web.pdf`

International visitor arrivals to New Zealand: September 2015 (Overseas visitor arrivals to New Zealand by country of residence and selected characteristics). (2015). Statistics New Zealand Tatauranga Aotearoa. Available from `http://www.stats.govt.nz/browse\_for\_stats/population/Migration/`

`international-visitor-arrivals-sep-15.aspx`

Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1193.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. *Talker Variability in Speech Processing*, 145–165.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, 505 - 528.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384.

Juskan, M. (2016). *Production and perception of local variants in Liverpool English: Change, salience, exemplar priming*. Unpublished doctoral dissertation, University of Freiburg, English Department.

Karpinska, M., Uchida, S., & Grenon, I. (2015). Vowel perception by listeners from different English dialects. In *The 18th international congress of phonetic sciences.*

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *The Psychology of Learning and Motivation*, *58*, 117–165.

Kendall, T., & Fridland, V. (2012). Variation in perception and production of mid front vowels in the US Southern Vowel Shift. *Journal of Phonetics*, *40*(2), 289–306.

Kendall, T., & Thomas, E. R. (2009). Vowels: Vowel manipulation, normalization, and plotting in R. *R package, version*, *1*.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Koops, C. (2010). /u/-fronting is not monolithic: Two types of fronted/u/in Houston Anglos.

University of Pennsylvania Working Papers in Linguistics, *16*(2), 14.

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, *14*(2), 12.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Attention, Perception, & Psychophysics*, *50*(2), 93–107.

Kuhn, R., Junqua, J.-C., Nguyen, P., & Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *Speech and Audio Processing, IEEE Transactions on*, *8*(6), 695–707.

Labov, W. (1963). The social motivation of a sound change. *Word*, *19*(3), 273–309.

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Labov, W., Yaeger, M., & Steiner, R. (1972). *A quantitative study of sound change in progress* (Vol. 1). US Regional Survey.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Langstrof, C. (2009). On the role of vowel duration in the New Zealand English front vowel shift. *Language Variation and Change*, *21*(3), 437.

Lawrence, D. (2015). Limited evidence for social priming in the perception of the bath and strut vowels. *Proceedings of the 18th International Congress of Phonetic Sciences.*, 244.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, *9*(2), 171–185.

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

dialectal speech recognition. In *Interspeech* (pp. 1458–1462).

Levi, S. V. (2014). Individual differences in learning talker categories: The role of working memory. *Phonetica*, *71*(3), 201–226.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, *130*(6), 4053–4062.

Liao, H., McDermott, E., & Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In *IEEE workshop on automatic speech recognition and understanding* (pp. 368–373).

Lincoln, M., Cox, S., & Ringland, S. (1998). A comparison of two unsupervised approaches to accent identification.

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. Psychology Press.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/and/l/. ii: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Lockrey, M. (2015, July). YouTube automatic craptions score an incredible 95% accuracy rate! *medium.com*. Available from `https://medium.com/@mlockrey/youtube-s-incredible-95-accuracy-rate-on-auto-generated-captions-b059924765d5#.xf73lddee` ([Online; posted 25-July-2015])

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886.

Lyle, S. (2008). *Dialect variation in stop consonant voicing*. Unpublished doctoral disserta-

tion, The Ohio State University.

Maclagan, M., & Hay, J. (2007). Getting fed up with our feet: Contrast maintenance and the New Zealand English "short" front vowel shift. *Language Variation and Change*, *19*(01), 1–25.

Mazzoni, D., & Dannenberg, R. (2000). Audacity (software). *The Audacity Team, Pittsburg, PA, USA*.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., et al. (2005). The AMI meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research* (Vol. 88).

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech*, 0023830914565191.

Meers, J. (2009). *The acquisition of front rounded and nasalized vowels of French by native speakers of English*. Unpublished doctoral dissertation, University of Calgary.

Meier, P., & Muller, S. M. (1998). IDEA: International dialects of English archive. *Accessed May*, *17*, 2005.

Milanesi, C. (2016, June). *Voice assistant anyone? yes please, but not in public!* Available from `http://creativestrategies.com/voice-assistant-anyone-yes-please-but-not-in-public/`

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329.

Milroy, L. (1980). *Language and social networks*. Blackwell Oxford.

Milroy, L., & Muysken, P. (1995). *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Morrison, G. S. (2013). Theories of vowel inherent spectral change. In *Vowel inherent spectral change* (pp. 31–47). Springer.

Morrison, G. S., & Assmann, P. F. (2012). *Vowel inherent spectral change*. Springer Science & Business Media.

Morrison, L. (2017, Jan). Speech analysis could now land you a promotion. *BBC News*. Available from `http://www.bbc.com/capital/story/20170108-speech-analysis-could-now-land-you-a-promotion`

Najafian, M., DeMarco, A., Cox, S. J., & Russell, M. J. (2014). Unsupervised model selection for recognition of regional accented speech. In *Interspeech* (pp. 2967–2971).

Najafian, M., Safavi, S., Hanani, A., & Russell, M. (2014). Acoustic model selection using limited data for accent robust speech recognition. In *Proceedings of the 22nd european signal processing conference* (pp. 1786–1790).

Nallasamy, U. (2016). *Adaptation techniques to improve ASR performance on accented speakers*. Unpublished doctoral dissertation, Carnegie Mellon University.

Nearey, T. M. (1978). *Phonetic feature systems for vowels* (Vol. 77). Indiana University Linguistics Club.

Nguyen, D.-P. (2017). *Text as social and cultural data: a computational perspective on variation in text*. Unpublished doctoral dissertation, University of Twente.

Nguyen, N., Shaw, J. A., Pinkus, R. T., & Best, C. T. (2016). Intergroup dynamics in speech perception: Interaction among experience, attitudes and expectations. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 16.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.

Nycz, J. (2013). Changing words or changing rules? second dialect acquisition and phonological representation. *Journal of Pragmatics*, *52*, 49–62.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Oder, A. L., Clopper, C. G., & Ferguson, S. H. (2013). Effects of dialect on vowel acoustics and intelligibility. *Journal of the International Phonetic Association*, *43*(01), 23–35.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing* (pp. 5206–5210).

Peñalosa, F. (1980). Chicano sociolinguistics: A brief introduction.

Pépiot, E. (2014). Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in parisian french and american english speakers. In *Speech prosody 7* (pp. 305–309).

Pierrehumbert, J. (2001). Lenition and contrast. *Frequency and the Emergence of Linguistic Structure*, *45*, 137.

Plunkett, L. (2010, September). *Report: Kinect Doesn't Speak Spanish (It Speaks Mexican).* online. Available from `http://kotaku.com/5627036/report-kinect-doesnt-speak-spanish-it-speaks-mexican`

Preston, D. R. (2002). Perceptual dialectology: Aims, methods, findings. *Trends in Linguistics Studies and Monographs*, *137*, 57–104.

Proença, J., Celorico, D., Lopes, C., Candeias, S., & Perdigão, F. (2016). Automatic annotation of disfluent speech in children's reading tasks. In *Advances in speech and language technologies for iberian languages* (pp. 172– 181).

Rabiner, L., & Juang, B.-H. (1993). Fundamentals of speech recognition.

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology*, *8*.

Rickford, J. R. (1999). *African American Vernacular English: Features, evolution, educa-*

tional implications. Wiley-Blackwell.

Ryan, C. (2013). Language use in the United States: 2011. *American community survey reports*, *22*, 1–16.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218.

Sawalha, M., & Abu Shariah, M. (2013). The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In *Proceedings of the 2nd workshop of arabic corpus linguistics.*

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., et al. (2010). Google search by voice: A case study. In *Advances in speech recognition* (pp. 61–90). Springer.

Scharenborg, O. (2005). Parallels between HSR and ASR: How ASR can contribute to HSR. In *Interspeech 2005-eurospeech 9th european conference on speech communication and technology* (pp. 1237–1240).

Scharinger, M., Monahan, P. J., & Idsardi, W. J. (2011). You had me at "hello": Rapid extraction of dialect information from spoken words. *Neuroimage*, *56*(4), 2329–2338.

Shahani, A. (2015, March). Now algorithms are deciding whom to hire, based on voice. *All Tech Considered: Tech, culture and connection.* Available from `http://www.npr.org/sections/alltechconsidered/2015/03/23/394827451/now-algorithms-are-deciding-whom-to-hire-based-on-voice`

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic

variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306–3316.

Smit, P. (2010). A review of eigenvoice adaptation.

Soto, V., Siohan, O., Elfeky, M., & Moreno, P. (2016). Selection and combination of hypotheses for dialectal speech recognition. In *IEEE international conference on acoustics, speech and signal processing* (pp. 5845–5849).

Squires, L. (2013). It don't go both ways: Limited bidirectionality in sociolinguistic perception. *Journal of Sociolinguistics*, *17*(2), 200–237.

Stoet, G. (2010). Psytoolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104.

Strand, E. A. (2000). *Gender stereotype effects in speech processing.* Unpublished doctoral dissertation, The Ohio State University.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, *60*(4), 487–501.

Szakay, A., Babel, M., & King, J. (2016). Social categories are shared across bilinguals' lexicons. *Journal of Phonetics*, *59*, 92–109.

Tatman, R. (2016). "i'ma spawts guay": Comparing the use of sociophonetic variables in speech and Twitter. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 18.

Telaar, D., & Fuhs, M. C. (2013). Accent-and speaker-specific polyphone decision trees for non-native speech recognition. In *Interspeech* (pp. 3313–3316).

Tjalve, M., & Huckvale, M. (2005). Pronunciation variation modelling using accent features.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Computer vision and pattern recognition* (pp. 1521–1528).

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during

online speech perception. *Language and Cognitive Processes*, *27*(7-8), 979–1001.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in society*, *1*(02), 179–195.

Turner, J. C., & Oakes, P. J. (1986). The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, *25*(3), 237–252.

Vaux, B., & Golder, S. (2003). The Harvard dialect survey. *Cambridge, MA: Harvard University Linguistics Department*.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition*, *110*(2), 254–259.

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, *2*(1), 219–237.

Wang, N. J., Lee, L.-S., Seide, F., & Lee, L.-S. (2001). Rapid speaker adaptation using a priori knowledge by eigenspace analysis of MLLR parameters. In *International conference on acoustics, speech, and signal processing* (Vol. 1, pp. 345–348).

Wang, Y.-Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *Automatic speech recognition and understanding* (pp. 577–582).

Warren, P., & Hay, J. (2012). Methods and experimental design for studying sociophonetic variation. In A. Cohn, C. Fougeron, & M. Huffman (Eds.), *The oxford handbook of laboratory phonology* (p. 634-642). Oxford University Press.

Watson, C. (2014). Mappings between vocal tract area functions, vocal tract resonances and speech formants for multiple speakers. In *Fifteenth annual conference of the international speech communication association.*

Watson, C. I., Maclagan, M., & Harrington, J. (2000). Acoustic evidence for vowel change

in New Zealand English. *Language Variation and Change*, *12*(01), 51–68.

Wells, J. C. (1982). *Accents of English* (Vol. 1). Cambridge University Press.

Wheatley, B., & Picone, J. (1991). Voice across America: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, *1*(2), 45–63.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372–384.

William, F., Sangwan, A., & Hansen, J. H. (2013). Automatic accent assessment using phonetic mismatch and human perception. *Audio, Speech, and Language Processing, IEEE Transactions on*, *21*(9), 1818–1829.

Wolfram, W., & Schilling, N. (2015). *American English: dialects and variation* (Vol. 25). John Wiley & Sons.

Wright, R. (2004). A review of perceptual cues and cue robustness. *Phonetically Based Phonology*, 34–57.

Wu, K., & Childers, D. G. (1991). Gender recognition from speech. part i: Coarse analysis. *The Journal of the Acoustical Society of America*, *90*(4), 1828–1840.

Xiao, Y. J., Coppin, G., & Van Bavel, J. J. (2016). Perceiving the world through group-colored glasses: A perceptual model of intergroup relations. *Psychological Inquiry*, *27*(4), 255–274.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., et al. (2016). The Microsoft 2016 Conversational Speech Recognition System. *arXiv preprint arXiv:1609.03528*.

Ye, G., Liu, C., & Gong, Y. (2016). Geo-location dependent deep neural network acoustic model for speech recognition. In *IEEE international conference on acoustics, speech and signal processing* (pp. 5870–5874).

Yilmaz, E., Pelemans, J., et al. (2014). Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model. In *Proceedings interspeech 2014* (pp. 969– 972).

Young, S., Woodland, P., & Byrne, W. (1993). *Htk: Hidden markov model toolkit v1. 5.* Cambridge Univ. Eng. Dept. Speech Group and Entropic Research Lab. Inc., Washington DC.

Young, S. J., Odell, J. J., & Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on human language technology* (pp. 307–312).

Zanten, E. v., & Heuven, V. van. (1984). A cross dialect study of vowel perception in Standard Indonesian. *Proceedings of the 10th International Congress of Phonetic Sciences*, 634 - 640.

Appendix

Data for YouTube automatic caption comparison (Accent tag)

Table 8.1: Data used in analysis. "Correct" indicates the number of words in the word list portion that were correctly transcribed by the automatic captions, "Total" that number of words that that talker's word list included. Differences in the total are due to variations in the version of the accent tag, as discussed above.

State	Gender	Youtube URL	Correct	Total	Year
California	F	NsJqya8XLl4s	30	43	2011
California	M	YQR7M5oTPNg	22	28	2011
California	M	N4UwEOKuiRo	18	28	2011
California	M	eTalf-BeTM8	24	28	2011
California	M	4slOxqd532s	24	28	2011
California	F	wnqo4qoYySo	18	29	2011
California	F	otg1nTM0_H8	16	43	2011
California	F	kwQU3vDmgYQ	23	38	2012
California	F	PtSoXSRS8UU	19	31	2013
California	M	VNsvkysVFas	21	30	2013
California	M	5wgm-EZqGAA	14	32	2013
California	F	mWgyIbIjVnc	12	30	2013
California	F	fIUCdo12X5w	21	30	2013
California	M	WqKQPv79UJI	18	30	2014

California	M	utyLAb4j5As	15	27	2015
California	F	62D9lDWGPl0	23	33	2015
Georgia	M	83ECQRvu-j0	24	29	2011
Georgia	F	Bws26R40kZ4	8	28	2011
Georgia	M	j0Jp0yfUNPo	13	29	2012
Georgia	F	neKnBSKaHTY	12	32	2012
Georgia	M	9HnWf96h9Cg	16	28	2012
Georgia	F	9blh3pUN4G8	12	31	2012
Georgia	M	MBwH2Uqn-6Q	9	28	2012
Georgia	M	lvfRhHXL2u8	13	28	2012
Georgia	F	9V2obSQc9kk	18	28	2012
Georgia	M	HDlmJlhhIBM	22	31	2013
Georgia	F	Of7BkQgJ7jw	21	28	2013
Georgia	F	3yuPxbs79uI	11	23	2013
Georgia	F	fGxsH6SgKYk	20	28	2014
Georgia	M	oKIIRLbRhwY	15	27	2015
Georgia	M	vk-7CMpT38E	16	28	2015
Georgia	F	vk-7CMpT38E	12	28	2015
NewEngland	M	G2tQaunibnU	18	29	2011
NewEngland	M	sEFC9IG6pyc	17	28	2011
NewEngland	F	bHXZd7wn0hE	24	45	2011
NewEngland	M	p7MgHBBkCGM	26	28	2012
NewEngland	F	5VVgx1WZm6I	16	30	2012
NewEngland	F	gCwvjjkdCpk	9	28	2012

NewEngland	F	5VVgx1WZm6I	14	28	2012
NewEngland	M	xEyU1mvEQtE	19	27	2012
NewEngland	F	PeI8HaAONqk	4	27	2012
NewEngland	M	XLiBvyERp4Y	18	28	2012
NewEngland	M	id2zwVSUXvQ	26	28	2012
NewEngland	F	vF9-PMMGxzA	7	28	2012
NewEngland	F	US4V4OZXK3M	17	31	2013
NewEngland	M	nhZN4-ue3m4	17	27	2015
NewEngland	F	5Wcm9pG-ckw	22	30	2016
NewEngland	M	5GqGw6eA_TI	16	31	2016
NewZealand	F	t-Rup_fd-1w	24	43	2011
NewZealand	M	liieen2DgAY	20	27	2011
NewZealand	M	4yTGJ2vOqnA	29	31	2012
NewZealand	F	xKZxB711t-c	8	28	2012
NewZealand	M	kfWwGMHAmd0	39	53	2012
NewZealand	F	BoSJk9K0ePc	11	28	2013
NewZealand	M	jsZDWmeR-Kg	21	28	2013
NewZealand	M	TWy41cJY3FI	24	33	2013
NewZealand	M	4_SeIK3JZxo	19	31	2013
NewZealand	F	4nmyWC_jIwc	15	38	2013
NewZealand	F	ygB-FBd-zsk	9	31	2014
NewZealand	F	_oX02QmYfYU	12	28	2014
NewZealand	F	Uj76F2eYI0E	15	28	2014
NewZealand	M	3rMrdzpfrHk	19	31	2016

NewZealand	M	qICxo8RCm-s	17	28	2016
NewZealand	F	Amb9mhW3EXQ	6	31	2016
Scotland	F	MfO8F6Uz_qM	16	28	2011
Scotland	M	3KYfWKHS9xw	11	28	2011
Scotland	M	s_5gKgIb2PY	8	27	2012
Scotland	F	JatocJtUQRM	12	30	2012
Scotland	M	W_Jskhofo5Q	15	31	2012
Scotland	F	eKBHiAvxArw	6	31	2012
Scotland	M	6txxRUnWG-0	7	39	2012
Scotland	M	r03gVBZZISM	6	28	2012
Scotland	F	r9hluC75MX0	17	31	2013
Scotland	M	extq6S6WZQU	18	37	2013
Scotland	F	gnm6W4WELR4	14	30	2013
Scotland	F	T5CAXmNqXaM	14	28	2014
Scotland	M	QyOZTqTIF6s	16	31	2015
Scotland	M	3X0yXlWdF1M	31	55	2015
Scotland	F	t-JmhI9fb2U	11	35	2016
Scotland	F	k3rbB4ceUtM	13	26	2016

Table of talkers whose speech was used in evaluation in section 3.2.1

Words used to elicit New Zealand vowels

(The pre-rhotic vowels were not used in the experiment, and tokens for this experiment were only taken starred environments.)

1. *had

Filename	**States**	**Gender**	**Age**	**Born**	**Ethnicity**
alabama4	AL	female	70	1928	African-American
alabama8	AL	female	67	1934	African-American
alabama9	AL	female	60	1942	African-American
alabama12	AL	male	20	1980	African-American
alabama13	AL	male	24	1983	African-American
michigan9	MI	male	20	1984	African-American
michigan15	MI	male	22	1993	African-American
alabama11	AL	male	20	unknown	African-American
michigan4	MI	female	unknown	1948	Caucasian
alabama3	AL	female	50	1949	Caucasian
michigan2	MI	male	50	1950	Caucasian
michigan14	MI	female	62	1952	Caucasian
michigan6	MI	female	unknown	1955	Caucasian
michigan3	MI	male	43	1957	Caucasian
california6	CA	male	52	1958	Caucasian
michigan5	MI	female	unknown	1965	Caucasian
michigan12	MI	male	36	1972	Caucasian
michigan1	MI	male	24	1976	Caucasian
michigan7	MI	female	unknown	1976	Caucasian
michigan8	MI	male	27	1978	Caucasian
alabama1	AL	male	21	1979	Caucasian
alabama2	AL	female	20	1980	Caucasian
alabama5	AL	male	unknown	1980	Caucasian
alabama6	AL	male	unknown	1981	Caucasian
california2	CA	female	21	1981	Caucasian
california4	CA	female	24	1981	Caucasian
michigan10	MI	female	20	1985	Caucasian
alabama7	AL	male	13	1988	Caucasian
california10	CA	male	20	1996	Caucasian
alabama10	AL	female	unknown	unknown	Caucasian
california1	CA	female	21	unknown	Caucasian
michigan11	MI	male	27	1979	Caucasian
michigan13	MI	male	40	1968	Mixed
california5	CA	male	37	1969	Mixed
california7	CA	female	36	1978	Native-American
california8	CA	male	26	1988	Mixed
california9	CA	female	16	1998	Mixed
kryker	GenAm	female	unknown	1946	unknown
california3	CA	male	unknown	unknown	unknown
ahager	GenAm	female	unknown	unknown	unknown
bpope	GenAm	male	unknown	unknown	unknown
dhague	GenAm	female	unknown	unknown	unknown
earmstrong	GenAm	male	unknown	unknown	unknown
jgoldes	GenAm	male	unknown	unknown	unknown
jjohnson	GenAm	male	unknown	unknown	unknown
kscott	GenAm	female	unknown	unknown	unknown
mdwyer	GenAm	female	unknown	unknown	unknown
pmeierGA	GenAm	male	unknown	unknown	unknown
rcook	GenAm	female	unknown	unknown	unknown
rfrye	GenAm	female	unknown	unknown	unknown

Table 8.2: Table of talkers whose speech was used in evaluation.

2. hard
3. *head
4. *heed
5. herd
6. hid
7. hoard
8. hod
9. hood
10. hud
11. who'd