Subphonemic Variation in English Stops: Studies using Automated Methods and Large-Scale Data

by

Lisa Lipani

(Under the Direction of Margaret E. L. Renwick)

Abstract

This dissertation deals broadly with subphonemic variation in English stops /p t k b d g/. Each chapter deals with a different aspect of these stops: Chapter 2 examines long-distance coarticulatory properties of these stops, and how there is a relationship between the patterns of voicing and the acoustic segments surrounding them (making distinctions beyond the phonological voiced/voiceless, including 5 different patterns of voicing based on acoustics). Using machine learning to classify the voicing pattern from the long-distance acoustics shows that there is a relationship between these patterns and the acoustics of the surrounding segments. Chapter 3 examines voice onset time in these stops in the Digital Archive of Southern Speech (Kretzschmar et al., 2013; Kretzschmar et al., 2019) and how this subphonemic variable is affected by several linguistic and sociolinguistic variables. Results show linguistic variables to be significant, both in line with and clarifying previous research. Finally, Chapter 4 improves the alignment results of the Montreal Forced Aligner by adding pronunciation variants featuring t/d deletion, and this technique will prove to be a useful tool in future sociolinguistic research. Additionally, this chapter examines the subphonemic properties of /t d/ in an effort to understand the acoustic correlates that influence MFA's determination. For /d/, the burst was found to influence MFA's decision, and for /t/, the closure, burst, and other variants were found to influence MFA's decision.

Index words: [Phonetics, Stop Consonants, Subphonemic Variation, Reduction, Coarticulation, Voice Onset Time, Forced Alignment, t/d Deletion, Large-scale Data]

Subphonemic Variation in English Stops: Studies using Automated Methods and Large-Scale Data

by

Lisa Lipani

B.A., Augusta University, 2013

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

Doctor of Philosophy

Athens, Georgia

2021

Subphonemic Variation in English Stops: Studies using Automated Methods and Large-Scale Data

by

Lisa Lipani

Major Professor: Margaret E.L. Renwick

Committee: John T. Hale
William Hollingsworth
Keith Langston

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2021

Acknowledgments

First and foremost, I would like to extend my deepest gratitude to the chair of my committee, Peggy Renwick, who has truly gone above and beyond in supporting me and my research in this process. The completion of my dissertation would not have been possible without her invaluable guidance and advice. She has been and remains an important role model for me as a linguist. I am also grateful to rest of my committee: John Hale, for all his computational expertise and the opportunity to TA natural language processing; William Hollingsworth, for encouraging my interest in speech technology, and Keith Langston, for all his unwavering support since my first day at UGA.

I would also like to thank the Linguistics Department at UGA for funding my graduate studies and for the opportunities to teach; the UGA Graduate School for the summer dissertation completion fellowship, which allowed me to focus on my dissertation research and complete my second chapter; and the National Science Foundation and American Dialect Society, whose support funded my research in the Linguistic Atlas Project.

I am also grateful to my undergraduate linguistics teachers, especially Joe Kuhl and Chris Botero, to whom I owe my interest in the field and my foundational knowledge of linguistics.

I want to thank my family for all their unconditional love and constant support: my mom, Marcia Lipani; my dad, Donald Lipani; and my sister, Sara Lipani. I would also like to acknowledge my animals: my dogs, Cricket and George, and my donkeys, Daisy, Maggie, Marianne, and "Monkey", for their companionship.

Finally, I'd like to thank my fellow graduate students Lindsey Antonini, Mike Olsen, Rachel Olsen, Trevor Ramsey, Paula Rawlins, and Joey Stanley for their friendship, advice, encouragement, and collaboration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

This dissertation deals broadly with subphonemic variation in speech. Subphonemic variation is also known as phonetic detail, i.e., detail beneath the level of the phoneme, which is an overlooked but important area of study. Research concerned with phonetic detail has implications for many subfields of linguistics and holds the key to a more granular understanding of systematic variation that can reveal linguistic structure, probabilistic factors of language, and what makes speech sound natural.

This dissertation consists of three separate, but related, studies. First, Chapter 2 shows that there is a relationship between subphonemic variation present in the acoustics surrounding a segment and that segment's phonetic voicing properties. Second, Chapter 3 shows that several linguistic variables affect the subphonemic property of voice onset time in Southern speech. Finally, Chapter 4 shows that modifying the dictionary of the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) can improve its ability to detect variation in phonetic detail in terms of word-final t/d deletion, and additionally, that certain subphonemic properties of the stops /t d/ influence MFA's determination of the presence of /t/ or /d/.

This current chapter provides an overview of phonetic detail in section 1.2, introduces the research questions this dissertation answers in section 1.3, and provides a map of the chapters that follow in section 1.4.

1.2 Phonetic Detail

The following section provides a definition and brief overview of phonetic detail/subphonemic variation, including an overview of consonant variation (subsection 1.2.1), followed by three phenomena that result in this subphonemic variation: coarticulation (subsection 1.2.2), sociolinguistic variation (subsection 1.2.3), and reduction (subsection 1.2.4). This section then contains an explanation of why phonetic detail is critical to phonology (subsection 1.2.5), and provides some concluding remarks on phonetic detail (subsection 1.2.6).

In phonology, there has been a traditional focus on the phoneme — a discrete unit of speech in a mental representation. However, there is systematic variation beneath the level of the phoneme that can reveal linguistic structure and probabilistic factors of language. Additionally, this subphonemic variation is said to contribute to perceptual coherence (Hawkins & Smith, 2001), which is what makes speech sound "natural".

The definition of the term phonetic detail, according to Hawkins and Smith, is "anything that is not considered a major, usually local, perceptual cue for phonetic contrasts in the citation forms of lexical items"[1] (2001, p. 479); this includes subphonemic variation, i.e., variation beneath the level of the phoneme, but also variation at the level of the phone, such as complete assimilation (e.g., the phoneme /n/ in *green beans* being realized as [m] in *gree*[m] *beans* as a result of the phenomena of coronal place assimilation), as this is not a phonetic cue for lexical contrast.

[1]One area of research this dissertation is concerned with is voice onset time. While voice onset time is a perceptual cue for word-initial voicing, I maintain that this variation can still be considered under the umbrella of phonetic detail, as it is beneath the level of a phoneme, and additionally, is not a perceptual cue in all word positions, e.g., word-finally.

Many studies have investigated phonetic detail and its relationship to linguistic structure in some way. For example, Lavoie (2002a) examined the detail of the words *for* and *four* and found that segmental context, stress, pitch accent, and boundary affected the realization of *for* and *four* differently. Jurafsky, Bell, and Girand (2002) found the infinitive *to* to be shorter and more reduced than the preposition *to*. Furthermore, Hawkins and Smith (2001) identified differences in the morpheme {mis-} in the words *mistimes* and *mistakes* that was due to the status of {mis-} as a true prefix (in *mistimes*) and {mis-} as a pseudo prefix (in *mistakes*). When {mis-} is a true prefix, durational differences in the phonetic segments making up the morpheme lead to it being perceived with a "heavier beat" than the {mis-} that is a pseudo prefix. Local (2003) found that /am/ of *lime* and *I'm* differ in their phonetic detail of its labiality and nasality because of the function of these words, as *I'm* is a function word while *lime* is a content word. Additionally, Local (2007) found that *so* and *anyway* are realized differently in terms of loudness, pitch, and other phonetic parameters based on communicative function. Crucially, this variation in phonetic detail is systematic, not random, and it reflects something about the linguistic structure, communicative situation, or context (probabilistic factors, frequency, predictability, etc.); according to Hawkins this "[d]etailed phonetic form varies systematically with linguistic category and communicative function" (2012, p. 162).

The next subsections outline the processes that give rise to this fine phonetic variation; the first of these discusses the ways in which consonants vary in general (subsection 1.2.1).

1.2.1 Consonant Variation

A great deal of study of variation in phonetic detail has dealt with vowels; however, as Lavoie (2002b) points out, consonants can exhibit variation from their citation form as well, especially in spontaneous speech. This section discusses several well-known phenomena that lead to variation in English consonants, focusing on, but not limited to, stops. This section is

organized partially by place of articulation: first, velar consonants are discussed; then coronal consonants; followed by labial consonants. Finally, this section discusses consonant variation not unique to place. In order to provide some context for this dissertation's focus on stops, this section also discusses subphonemic variation of the acoustic components of stops.

It is well known that English velar consonants can vary in manner and place of articulation. An investigation of English /k/ from the MIT Map Task recordings shows that this consonant can vary in manner of articulation; it can be realized as an incomplete stop with frication, with voicing, as an approximant, or as a glottal fricative, as shown in Figure 1.1.

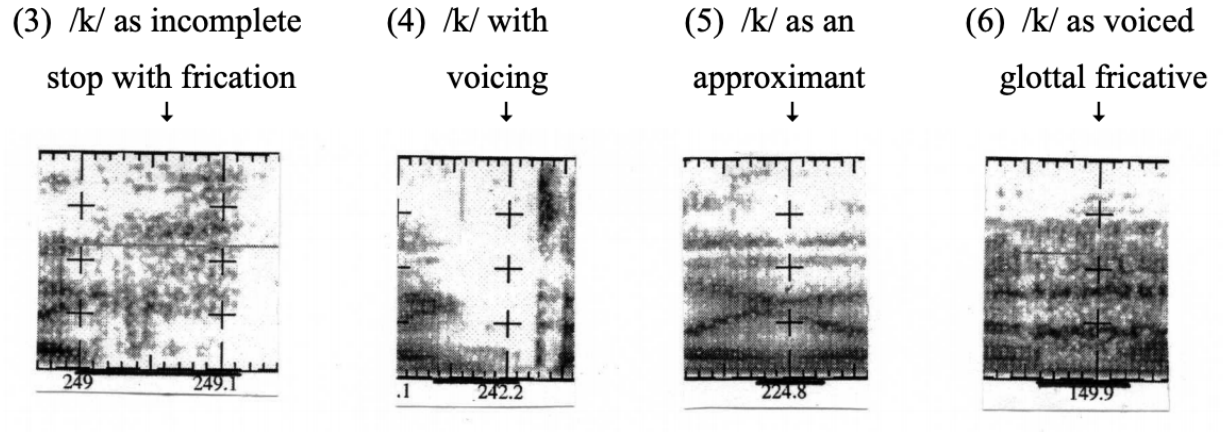

Figure 1.1: /k/ variation in MIT Map Task recordings, reproduced from Lavoie (2002b, p. 41)

A comparison to Spanish /k/ revealed that this variation is constrained by the phoneme inventory, as Spanish /k/ did not show variation similar to English /k/ because the resulting sound would be too similar to Spanish /x/ (Lavoie, 2002b). In addition to variation in *manner* that /k/ can exhibit, English velar stops /k, g/ are also well known to vary in *place*. For example, a well-attested phenomena in many introductory linguistics or phonetics textbooks is the difference in *key* [k$^{\text{h}}_{+}$i] and *coo*, [k$^{\text{h}}$u], where /k/ is fronted due to the influence from the high front vowel /i/[2]. This same process happens with /k/'s voiced counterpart /g/, as seen in *geese* [g$\underset{+}{\text{i}}$s] compared to *goose* [gus].

[2]However, it is important to point out that "fronting of velars is a gradient effect, less extreme than phonemic palatalization of velars" (Keating & Lahiri, 1993, p. 73).

Coronal stop consonants are also known to vary from citation form a great deal in English. It is well-known, for example, that word-initial stops feature aspiration (e.g., *top*, [tʰɑp]), word-medial stops may be reduced in certain phonetic contexts (e.g., *butter*, [ˈbʌɾəɹ]), and word-final stops may be unreleased (e.g., *apt*, [æpt̚]). As discussed in Chapter 4, the coronal consonants /t/ and /d/ in the Buckeye Corpus (Kiesling, Dilley, & Raymond, 2006) could be transcribed as [ð], [ɾ], [t̪], [d̪], [tʃ], or [dʒ], or /t/ and /d/ can be deleted entirely, which is especially likely in word-final consonant clusters, (e.g., *iced tea*, [aɪs ti]). Similarly to velar consonants, coronal consonants can be affected by their adjacent consonants. For example, Rogers (2014, p. 48) gives the pronunciation of *bird* as [bəɹd̪], as the /d/ becomes slightly retroflex in the presence of the /ɹ/.

Word-final coronal consonants in general (i.e., /t, d, n, s, z/) are observed to undergo place assimilation, as are labial consonants[3]. Place assimilation is a process in which these consonants change their place of articulation to match to the following consonant (e.g., *in Paris*, [ɪm pɛɹɪs]). Avery and Rice (1989) and Coleman, Renwick, and Temple (2016) show that labial and coronal nasal consonants /m, n/ undergo assimilation, for example, *from Kingston* realized as [fɹʌŋ kɪŋstən] (Avery & Rice, 1989) or *some cream* realized as [sʌŋ kɹim] (Coleman et al., 2016).

Some instances of variation can affect stops at all places of articulation, in terms of manner of articulation and even airstream mechanism. Riebold (2011) found that stops in general, but velars especially, could spirantize[4] in Pacific Northwestern English, and were more likely to do so than bilabials[5]. Further, there is evidence that stops could vary in airstream mechanism. Typically, English sounds feature a *pulmonic egressive* airstream, in which airflow is pushed *out* of the lungs. Standard English stop consonants feature an occlusion of the airflow,

[3]There is anecdotal evidence that velar oral stops also assimilate in place, for example, *like that* realized as [laɪt ðæt] (Barry, 1985). However, I found no sufficient evidence to claim that velars can assimilate (Lipani, 2017).

[4]A process in which a stop becomes a fricative.

[5]Coronal stops were not included in this study due to their highly variable nature, as previously discussed.

followed by a release of air; the air is moving from the lungs outward. During the occlusion of an implosive stop, however, the glottis moves downward, creating negative air pressure (i.e., rarefaction) in the oral cavity, and then the stop is released as air exits the lungs; implosive stops are made with both a glottalic ingressive and pulmonic egressive airstream mechanism. While English stops are usually described as pulmonic egressive, Jacewicz, Fox, and Lyle (2009) describes an "implosive characteristic" to the release of Southern American English bilabial stops. In an attempt to quantify this observation, Husain (2017) used acoustic analysis to study North Carolina speakers' production of voiced stops and found stops at all places of articulation can be somewhat implosive.

Subphonemic properties of stops can also vary in many ways, especially due to their manner of articulation as stops. Stops feature three distinct phases: the closure, burst, and aspiration (if applicable), which are discussed further in subsubsection 1.2.4.2. For example, in a study of /t/ in Dutch[6], Schuppler (2012) measured six components of /t/s, shown in Table 1.1[7].

The above examples give a brief overview of the great deal of variation present in English stop consonants. This dissertation focuses on three different processes that can give rise to the variation mentioned above: coarticulation (subsection 1.2.2), sociolinguistic variation (subsection 1.2.3), and reduction (subsection 1.2.4). Within each of these larger topics, a specific phenomenon is explored in each chapter in turn: coarticulation of phonetic voicing and long-distance acoustics, the sociophonetic variation of the subphonemic cue voice onset time, and finally, the deletion (one end of the reduction continuum) of word-final /t d/.

[6]Schuppler (2012) is not studying English; however, Dutch and English are both Germanic languages, and the categorization of these subphonemic properties is applicable to the parts of stops in general, not specific to English.

[7]Chapter 4 will deal with these subphonemic properties in more detail.

[8]This was considered a subphonemic property of /t/ because Schuppler "observed that acoustically reduced /t/s abruptly followed by fiction of the next segment are perceived as strong /t/s (compared to /t/s where the frication of the next segment starts smoothly)" (2012, p. 9).

Table 1.1: Subphonemic Properties in Schuppler (2012)

Variable	**Levels**
Closure	Present Realized with frication Realized with nasal frication Realized with nasal murmur
Voicing of the constriction	Voiced Unvoiced
Burst	One Multiple None
Strength of the burst	Strong Weak
Frication of the following segment[8]	Started gradually during /t/ Started abruptly during /t/ Started simultaneously with /t/ Absent from /t/ Started gradually during following consonant Started abruptly during following consonant Started simultaneously with following consonant Absent from following consonant
Frication place of articulation	Alveolar Not alveolar

1.2.2 Coarticulation

Coarticulation is the process during which one speech sound affects neighboring speech sounds (Daniloff & Hammarberg, 1973). Coarticulation can further be defined as *carryover*, or "left-to-right, perservative", where a preceding segment affects a following segment. Conversely, there is also *anticipatory* coarticulation, or "right-to-left", where a following segment affects a preceding segment. In English, carryover coarticulation is thought to be more prevalent (Hoole, Nguyen-Trong, & Hardcastle, 1993). Furthermore, coarticulation can be local, where

a segment affects another segment directly next to it, or long-distance, where a segment affects another segment that is not directly next to it. Coarticulation can affect the acoustics of the surrounding sounds in subphonemic ways. For example, Öhman (1966) found that frequencies in the second vowel of VCV utterances of American English and Swedish speakers were dependent on both the preceding consonant and vowel, contradicting theories that stop consonant loci are static; the subphonemic detail varies with coarticulation.

A great deal of coarticulation studies deal with effects of a vowel on another vowel (i.e., vowel-to-vowel coarticulation, e.g., Öhman (1966)). However, there are studies dealing with how consonant change is affected by coarticulation. One example of this is the assimilatory process that can take place, often to word-final consonants. For example, in *line perfectly*, the /n/ of *line* takes on the labial characteristics of the following segment /p/ to become [m]. This example and an additional example of this coarticulatory process, assimilation, is given in Table 1.2.

Table 1.2: Assimilation examples from Gow (2002)

Orthographic form	**Phonemic form**	**Assimilated form**
line perfectly	/laɪn pɛɹfɛktli/	[laɪm pɛɹfɛktli]
mat belonged	/mæt bilɔŋd/	[mæp bilɔŋd/]

Another example of consonant variation as a result of coarticulation is shown in English /ɹ/, which exhibits affrication when preceded by a coronal stop /t/ or /d/, observed by Read (1971) and studied by Smith, Mielke, Magloughlin, and Wilbanks (2019). For example, *tree* and *dream* may be pronounced as [tʃɹi] and [dʒɹim] respectively (Smith et al., 2019, p. 3). Consonants can be the subject of change, but they can also affect coarticulatory change. For example, the English liquids /l/ and /ɹ/ are shown to have long distance resonances present in the rest of the utterance (Kelly & Local, 1989). This and other studies are discussed at length in subsection 2.1.2.

1.2.2.1 Factors Affecting Coarticulation

Several linguistic factors have been shown to influence the strength of coarticulatory processes, including prosody, stress, time, and frequency, each discussed in this section below.

Cho (2004) found that vowel-to-vowel coarticulation varies as a result of prosodic structure. For example, vowels that were located in sententially-stressed, domain-initial positions had weaker vowel-to-vowel coarticulatory effects; additionally, vowels were less likely to exhibit coarticulatory influence across higher prosodic boundaries. Additionally, Cho, Kim, and Kim (2017) found that prosodic position (domain-initial vs domain-final) influenced the degree of coarticulatory influence of a nasal on a preceding vowel.

Additionally, stress was found to influence coarticulation. De Jong, Beckman, and Edwards (1993) found that the jaw was lower in the syllable [pɑp] when it was nuclear-accented (a result that also suggests the influence of prosody). The authors draw on the relationship between hypoarticulation and hyperarticulation proposed by Lindblom (1990) and less-stressed and more-stressed syllables to conclude that stressed syllables are less likely to be subject to coarticulatory properties. Further, in a real- and apparent-time[9] sociolinguistic study of Philadelphia English, Zellou and Tamminga (2014) found that the nasality of vowels, present due to nasal-to-vowel coarticulation, changed over time, with older speakers more likely to exhibit nasality, middle-aged speakers less likely to exhibit nasality, and younger speakers more likely to exhibit nasality in their vowels.

Additionally, Zellou and Tamminga (2014) found an effect of lexical frequency; more frequent words were more likely to have nasalized vowels, i.e., have nasal-to-vowel coarticulation.

[9]A real-time study is one in which sound change is examined through longitudinal study of the same speakers; an apparent-time study is one that relies on the apparent-time hypothesis (Bailey, Wikle, Tillery, & Sand, 1991), which is the theory that the language one learns as a child is present in one's speech today. Therefore, apparent-time studies can look at the speech of different generations in order to sample speech from different periods of time.

1.2.3 Sociophonetic Variation

Sociophonetic variation, by definition, arises from social factors such as ethnicity, region, gender, and class. Labov, Yaeger, and Steiner's (1972) study on variation in American English vowels is considered to be the "foundations" of sociophonetics (Baranowski, 2013). Baranowski (2013) points out that until relatively recently, the term "sociophonetics" typically referred to the study of *vowel* variation, and a call for further study of consonants is made in Foulkes and Docherty (2006). However, the study of consonants within this field is expanding, and this dissertation makes a contribution to this expansion. As this dissertation focuses on consonants, so too will this section.

According to Thomas (2016), liquids (especially rhotics), and components of stops such as voice onset time, glottalization, aspiration, and voicing have all been the subject of consonantal sociophonetic research. A well-known study of sociophonetics of /ɹ/ is Labov's (1986) study in which social status was found to positively influence r-dropping. However, the sociolinguistic properties of /l/ and /ɹ/ are outside the scope of this dissertation (for a review, see Thomas (2016)), and the following paragraph will discuss previous sociophonetic research on *stops*. Stop consonants are somewhat different from other sounds in English in that they have three distinct components, presented here in order of occurence: 1. a complete closure of the articulators during which there is no (or minimal, if vocal folds are vibrating) acoustic energy, 2. a release burst present from the articulators opening once the air pressure behind them is great enough, and 3. an optional period of aspiration, a longer puff of air. The time between the burst of the stop and the onset of voicing is called voice onset time (VOT), and the presence of aspiration contributes to a longer VOT (Lisker & Abramson, 1964, 1967). VOT is a common way in which stops vary as a result of phonological voicing, language contact/bilingualism (e.g., Flege (2007), Newlin-Łukowicz (2014), among others), dialect, sex, and ethnicity (all discussed in detail in Chapter 3).

There is a call for the methodology of sociophonetics to change, and Chapter 4 of this dissertation is a small step in advancing these methods. Thomas (2016) points out that most sociophonetic studies rely on the auditory coding of trained linguists. Two issues that are briefly brought up in section 4.1 are 1. the fact that reliance on auditory coding is undesirable due to the perceptual ability to reconstruct reduced forms (Mitterer, 2011)[10] and 2. the time it would take to examine large-scale data manually is prohibitive. In response to the disadvantages of auditory coding, Thomas calls for the combination of "advanced phonetic methods and advanced sociological analyses" (2016, p. 1). Chapter 3 of this dissertation employs advanced phonetic methods, and Chapter 4 examines methodology that could be easily employed in order to automate advanced sociological acoustic phonetic research.

1.2.4 Phonetic Reduction

Reduction, a process that can affect phonetic detail, is a newer area (relatively speaking) in phonetic research. It has been defined as variants that are "characterized by incomplete articulatory gestures or fewer segments compared to the variants typical of read speech" (Ernestus & Warner, 2011, p. 254). Reduced speech can vary from citation form temporally, in the deletion of segments or shortening duration, or spectrally, by weakening processes, such as spirantization and centralization of formant values. Examples of reduction are given in Table 1.3, where the top two examples are reproduced from Ernestus and Warner (2011) and the bottom four examples are reproduced from Johnson (2004).

1.2.4.1 Factors Affecting Reduction

Lexical properties can affect reduction in terms of frequency and predictability. Frequency of a word plays a role in the likelihood of reduction occurring in English (Bell, Brenier, Gregory,

[10]This is not to dismiss auditory coding entirely. Chapter 4 of this dissertation involves manual coding that is sometimes based on auditory qualities of sound. Of course, I am only human, but as a trained phonetician, I place some trust in my judgments.

Table 1.3: Reduction examples from Ernestus and Warner (2011) and Johnson (2004)

Orthographic form	**Full phonetic form**	**Reduced phonetic form**
yesterday	[jɛstəɹdeɪ]	[jɛʃeɪ]
do you have time	[dujuhævtaɪm]	[djutɛm]
because if	[bikʰʌzɪf]	[kʰzɪf]
apparently (not)	[ʌpɛɹɛntli]	[pɛˠɾɪ]
hilarious	[hɪlɛriʌs]	[hlɛrɛs]
particular	[pʰɑrtɪkulər]	[pʰtʰɪkʰəˠ]

Girand, & Jurafsky, 2009; Davis & Cohn, 2020; Jurafsky, Bell, Gregory, & Raymond, 2001) and in Dutch (Pluymaekers, Ernestus, & Baayen, 2005), with more frequent words showing more reduction. Additionally, frequency effects were found in terms of the shortening of duration (Arnon & Cohen Priva, 2013; Gahl, Yao, & Johnson, 2012) and spectrally, in terms of a centralization of the vowel space (Munson & Solomon, 2004). More predictable words are more likely to be reduced (Bell et al., 2009; Mitterer & Russell, 2013), and similarly, words in collocations were likely to be reduced as well (Jurafsky, Bell, Fosler-Lussier, Girand, & Raymond, 1998). Related to frequency and predictability, informativity, which is the average predictability of a phone given previous phones in the word (Cohen Priva & Jurafsky, 2008), can influence phone duration, with more informative phones being longer and less subject to deletion (Cohen Priva, 2015; Cohen Priva & Gleason, 2020; Cohen Priva & Jurafsky, 2008). Additionally, Davis and Cohn (2020) shows that compositionality of a word can affect reduction processes, with more opaque compound words (e.g., *cupboard*) undergoing more reduction than transparent compound words (e.g., *blueberry*).

Of course, phonological factors play a role in reduction as well. As the number of phonemes increases, reduction increases, even when controlling for number of syllables (Mitterer, 2008). Mitterer (2008) suggests that the perceptual salience of speech sounds plays a role in reduction, with less perceptually salient sounds featuring higher rates of deletion. Additionally,

other phonetic properties such as phonological neighborhood density influence reduction. Phonological neighborhood density (sometimes abbreviated as PND) is the number of words that vary in only one segment (e.g., from deletion, insertion, substitution) from a given word. For example, the phonological neighborhoods of *cat* are *at*, *cats*, *cap*, etc. As phonological neighborhood density increases, so does the likelihood of reduction (Gahl et al., 2012; Yuan, Lin, & Liu, 2020).

Various aspects of discourse influence reduction. The mention of a word in discourse also affects reduction, as a word is more likely to be reduced on second or subsequent discourse mention (Fowler & Housum, 1987; Shockey, 2008). Furthermore, the style of conversation is thought to influence reduction, with more conversational speech more likely to be reduced than more formal speech in Dutch (Ernestus, 2000) and English (Fletcher, McAuliffe, Lansford, & Liss, 2015).

Additionally, sociolinguistic factors are shown to influence reduction, as Byrd (1994) found that male speakers were more likely to reduce than female speakers. Another example of this is that consonant cluster reduction in Tejano English shows differences from General American English[11] (Bayley, 1994).

As Clopper et al. (2018) point out, the interactions among these factors (e.g., the interaction between frequency and predictability found by Bell et al. (2009), the three-way interaction between probability, speech style, and prosody found by Baker and Bradlow (2009)) mean that the way in which these factors affect reduction is complex.

1.2.4.2 Reduction of Stops

As canonically defined, stop consonants consist of a complete closure, burst, and aspiration (if applicable), and so reduction can be examined on a subphonemic level based on these three properties. In a study that investigated the subphonemic properties of reduction, Schuppler (2012) carried out acoustic analysis that examined closure (present, realized with frication,

[11]This is discussed further in Chapter 4.

nasal frication, or nasal murmur caused by a preceding nasal), voicing (present or absent), and burst (present or absent, and further characterized by number of bursts, and whether they were strong or weak). In an extensive study of stop consonant variation, Warner and Tucker studied American English intervocalic stops /p k b g/ and the intervocalic flap [ɾ] (from /t d/) and found "greater reduction in more casual speech, greater reduction at the word boundary in higher frequency phrases but not internal to higher frequency words, and greater reduction between two unstressed syllables than after a stressed syllable (e.g., *limited* versus *status*)" (2011, p. 1615)

Reduction of stops also can be considered to include deletion. Schuppler (2012) and Coleman et al. (2016) outline two possibilities: 1. deletion is separate process, where a segment is categorically present or absent, and 2. deletion is one end of a gradient reduction spectrum. However, Raymond et al. (2002) argue that their finding of speech rate affecting word-internal /t d/ deletion in the Buckeye Corpus (Clopper & Pisoni, 2006) supports the idea that deletion is one end of a reduction continuum rather than a separate categorical process. Schuppler (2012) also came to the conclusion that deletion is an end of a reduction spectrum.

1.2.5 Phonetic Detail and Phonology

This dissertation is concerned with fine *phonetic* detail, and thus deals with the concrete measurements from the speech signal; however, these measurements are of interest to the phonologist as well. This section first explains the role of phonetic detail in phonological theories, and then moves on to discuss specific theories and approaches in more detail.

Hayes acknowledges the difference between phonetics and phonology, and the importance of phonetic issues to phonology:

> Phonetics studies speech sounds in ways that are close to the speech stream, focusing on production, acoustics, and perception and phonology, which tends to

> be more abstract, dealing not directly with the physical nature of speech sounds (**though that is of course quite relevant** [emphasis mine]), but rather with the largely unconscious **rules** [emphasis his] for sound patterning that are found in the mind/brain of a person who speakers a particular language (2011, p. 19).

A speaker's phonological knowledge is of course informed by phonetic realizations, but the mapping of the physical reality of the waveform to what is understood is a question central to phonology. Traditional phonological research assumed invariance in the acoustic signal, spurring on phonetic research that sought to find invariance in the acoustic signal. A well-known example of this kind of research is Delattre, Liberman, and Cooper (1955), a study in which researchers found formant transitions "pointed to" certain frequencies that correspond to places of articulation, as shown in Figure 1.2.

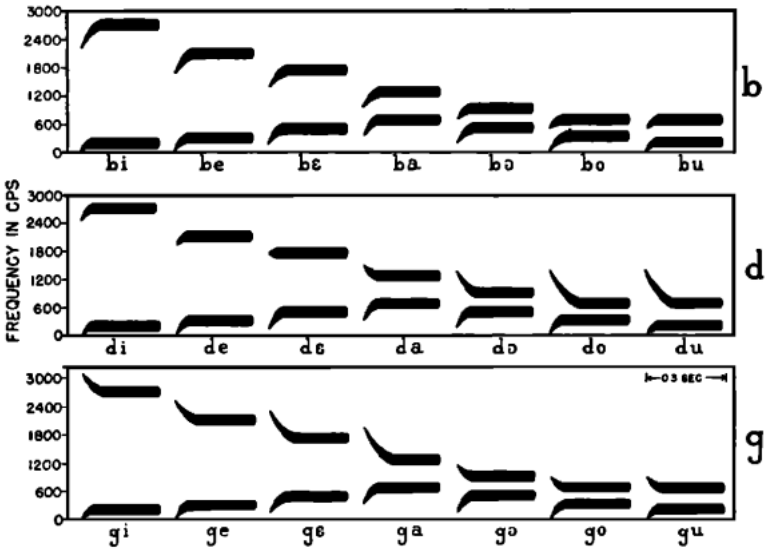

FIG. 1. Synthetic spectrograms showing second-formant transitions that produce the voiced stops before various vowels.

Figure 1.2: Formant transitions by place of articulation, reproduced from Delattre, Liberman, and Cooper (1955, p. 770)

However, there has not been a great deal of success in finding truly invariant properties of the acoustic signal (cf. Stevens and Blumstein (1981)). If this level of invariance is unlikely

to exist, the scope of "invariance" could be expanded to find what acoustic cues give rise to a meaningful contrast at the lexical level. For example, Coleman (2003) points out that phonetic differences *do* give rise to meaningful contrast: the phonetic realizations [lɛ̃ʔtʰ] and [lɛ̃nʔtʰ] could be taken to be representations of *lent*, while [lɛ̃ːnːt] could be a representation of *lend*. However, as discussed in previous sections, there is meaningful within-category variation that can occur at the level of the phone (e.g., *gree*[n] vs. *gree*[m] *beans*)[12] or subphonemically (e.g., a longer or shorter VOT based on socioindexical differences).

This focus on invariance at the level of the phoneme that is lexically meaningful has led researchers away from variation beneath the level of the phoneme that reflects some linguistic quality of the utterance; "over-emphasis on phonemes has deflected attention from the presence of other types of linguistic information in the speech signal and, in consequence, distorted the relative importance of various processes in models of how speech is understood" (Hawkins & Smith, 2001, p. 101). Hawkins (2003) provides a good example of the difference in the ways which *I don't know* can be pronounced, as shown in Figure 1.3.

Table 1
Ways to mean *I do not know* as an illustration of the centrality of phonetic fine detail to the understanding of full connotative meaning

1. *I…do…not…know*
2. *I do not know*
3. *I don't know*
4. *I dunno*
5. *dunno*
6. [ã̆ǎnːəŭ]
7. [ə̰ə̃ə̰]
8. but not [ə̰ə̃ə̰] or [m̩ m̩ m̩]

Versions 1–7 are all acceptable. Versions 3–5 are standard for normal relaxed interaction (if spoken without insolence). Versions 1 and 2 imply negative attitude (if spoken fluently and without absent-mindedness); version 6 is normally used between social equals, but can introduce a new opinion; version 7 is only possible when the participants are social equals and external contextual cues are very strong (intonation and stress are important, but have been omitted).

Figure 1.3: Pronunciations of *I don't know*, reproduced from Hawkins (2003, p. 375)

[12]Interestingly, when assimilation at the level of the phone occurs in lexically ambiguous contexts (i.e., *He picked up the* cone-comb *between his fingers*), Gow found that it does not create lexical ambiguity; "assimilation preserves acoustic evidence about the unmodified form of words" (2002, p. 174).

Thus, a central line of inquiry in the field of phonology is how listeners are able to understand speech despite its high level of variability (some of which is examined in this dissertation), i.e., how do listeners map sound to meaning, and how are phonological representations of lexical items stored in the mental lexicon? There are two major theories that address this question and address how fine phonetic detail either does or doesn't play a role in this process: Abstractionist Theories and Exemplar Theories [13].

1.2.5.1 Abstractionist Theories

In abstractionist models, all words are stored in the mental lexicon as a set of invariant abstract features (e.g., phonemes, phonological features), and surface representations that are actually pronounced undergo the application of phonological processes (phonological rules or constraints in Optimality Theory (Smolensky & Prince, 2004), as Ernestus (2014) points out) in order to arrive at what is stored in the mental lexicon. In abstractionist theories, fine phonetic variation of speech is *abstracted*[14] into phonological units for speech processing[15]. In other words, speech processing abstracts away all detail at an early stage.

A purely abstractionist model, one in which phonetic detail is totally forgotten, is disproven by evidence that fine phonetic detail arising from talker or situational effects affects speech processing (Cutler, Eisner, McQueen, & Norris, 2010). These abstractionist theories do not discount the existence of phonetic detail, but rather, claim that these variants are not stored in the mental lexicon, and there is some process in which the variation is abstracted into features that *are* stored in the lexicon. For example, Lahiri and Reetz (2002, 2010) propose a system in which acoustic cues are abstracted into binary feature values (e.g., [± strident]) that are then mapped onto the lexicon; this model is called the Featurally Underspecified Lexicon

[13] These two categories of theories also posit how speech production might occur, but what follows focuses on how variation is stored in the brain and perceived.

[14] Pierrehumbert (2016) points out that all scientific theory is abstract, though in this instance, I use the term *abstract* to refer to the idea that the representation of discrete words consists of an economical approach of minimally distinct representations of sounds.

[15] Auditory theories of speech perception and motor theories of speech perception typically start from the viewpoint that we perceive *abstract* categories as well.

(FUL) model. An example of this model's application is the perception of words that have undergone place assimilation (as discussed in subsection 1.2.1) where the feature [coronal] is underspecified and "listeners are insensitive to surface variations" (Nguyen, Wauquier, & Tuller, 2009, p. 6) that are present as a result of the assimilatory process.

Additional evidence for abstractionist models is the fact that listeners can be primed to hear certain phonemes based on lexical items. In McQueen, Cutler, and Norris's (2006) study, listeners were inhibited or primed to hear [s] or [f] based on a training period in which two groups heard an ambiguous token in a word-final position of words canonically ending in [f] or [s]. McQueen et al. argue that there is a "prelexical level of processing that codes abstract phonetic information (i.e., that somehow represents the category distinction between [f] and [s])" (2006, p. 1121). Casserly and Pisoni also point out that there is evidence for categorical perception, where "phoneme representations split potential acoustic continuums into discrete categories" (2010, p. 632). They give an example of voice onset time, a numeric variable that is a continuum, which is split into perceptual categories — voiced and voiceless.

Despite some evidence for abstraction, many of the fine phonetic detail effects previously discussed in this chapter are not accounted for in purely abstractionist models. In a landmark study, Goldinger (1998) found that listeners were likely to mimic the fine phonetic detail of an utterance that they heard. Furthermore, Pierrehumbert (2016) argues that frequency effects (such as those found by Jurafsky et al. (2002)) and the effect of indexical information[16] such as gender, age, dialect, social class, are stored in the memory. However, this is not to say that all abstractionist models dismiss fine phonetic detail. The difference in the treatment of fine phonetic detail between abstractionist theories and exemplar-based theories is best summarized by Nguyen et al.:

> Thus, the assumption that FPD [fine phonetic detail] has a role to play in speech perception is not specific to exemplar models and is also found in at least some

[16]Pierrehumbert uses indexicality to refer to information in the utterance about "the speaker, the social context, or the physical context" (2016, p. 41).

> abstractionist models. Exemplar models do diverge from abstractionist models, however, in assuming that in addition to being relevant to on-line speech perception and understanding, FPD is stored in long-term memory. More specifically, and as opposed to abstractionist models, exemplar models posit that lexical representations are phonetically rich (2009, p. 10).

Abstractionist models, even ones in which phonetic detail is attended to by the listener, are seemingly incompatible with the the fact that there is meaningful linguistic information in phonetic detail. In contrast to abstractionist theories, exemplar theories are another type of model that do acknowledge the role of phonetic detail in speech comprehension.

1.2.5.2 Exemplar Theories

Exemplar theory is a general theory about human classification in which individual examples (exemplars) are stored in the memory, and classification judgement about new instances is made based on these examples. An example illustrating how exemplar theory works is given below:

> People represent the category of "birds" by storing in memory the vast collection of different sparrows, robins, eagles, ostriches (and so forth) that they have experienced. If an object is sufficiently similar to some of these bird exemplars, then the person would tend to classify the object as a "bird". This exemplar view of categorization contrasts dramatically with major alternative approaches that assume that people form abstract summary representations of categories, such as rules or idealized prototypes (Nosofsky, 2011, p. 18).

In phonological terms, this means that speakers store the phonetic detail of sounds (rather than the idealized categories of abstractionist theories) in representations of words[17]

[17]Though most propose storage in terms of words, "[t]his, however, does not mean that sublexical units such as segments and syllabic constituents cannot have a psychological reality" (Nguyen et al., 2009, p. 11).

in the mental lexicon[18]. Then, the acoustics of a word to be categorized are matched to an exemplar of that word, i.e., "episodic traces in lexical representations" (Pierrehumbert, 2016, p. 40), containing its characteristics (e.g., syntactic and semantic properties). These detailed representations explain how people are able to have knowledge of fine phonetic detail (Pierrehumbert, 2001). The importance of fine phonetic detail to communication is as follows:

> Naturally, some of these factors can be analyzed into the well-accepted units of formal linguistic analysis: intonational phrases, feet, syllables, phonemes, and allophones associated with particular positioning in syllables of particular structure. But some of the important details are not readily accommodated by standard phonological-linguistic units; yet when they are systematically reflected in the speech signal, they, too, can be crucially important to communication (Hawkins, 2003, p. 374).

There is strong evidence for exemplar theory in terms of the variation previously discussed in this chapter and frequency effects. First, variation beyond segmental allophonic change supports exemplar-based theories, and furthermore, this variation is gradient (Guy, 2014). Additionally, the fact that higher frequency words are more likely to undergo reduction (as discussed in subsubsection 1.2.4.1) also supports exemplar theory. For example, a high-frequency word like *every* is more reduced than a medium frequency word like *memory*, which itself is more reduced than a low-frequency word like *mammary* (Hooper, 1976). Frequency (and predictability) effects are taken to be evidence for exemplar theory:

> Given a tendency for reduction during production, the phonetic representation of a word will gradually accrue more exemplars that are reduced, and these exemplars will become more likely to be chosen for production, where they may undergo further reduction, gradually moving the words of the language in a consistent

[18]Of course, a criticism of the idea that all phonetic detail is stored in the brain is called the "head-filling-up problem" (Neal Johnson, personal communication qtd. in Johnson (1997)).

> direction. The more frequent words will have more changes to undergo online reduction and thus will change more rapidly. The more predictable words (which are usually also the more frequent ones) will have a greater chance of having their reduced version chosen, given the context, and thus will advance the reductive change more rapidly (Bybee, 2002, p. 271).

Overall, the fact that 1. gradient subphonemic variation exists, 2. this variation reflects something about the linguistic structure (e.g., frequency), and 3. listeners can exploit this information to glean meaning is only adequately addressed by theories that include episodic memory traces. This is not to say that strictly exemplar-based theories are the only ones that are adequate, however.

1.2.5.3 Hybrid Models

There are advantages and disadvantages to both abstractionist and exemplar theories (e.g., failure to account for phonetic detail and head-filling-up problem, respectively), and more recent work suggests a hybrid approach is needed (Ernestus, 2014; Nguyen et al., 2009; Pierrehumbert, 2016). Indeed, Hawkins says, "It is clear we attend to details; it is clear we abstract to general categories: there seems no point polarising theories when the data support neither extreme" (2010, p. 482).

One conceptual approach[19] is Polysp (POLYsystemic SPeech understanding), proposed by Hawkins and Smith (2001). Firthian Prosodic Analysis, the basis for this model, is not well-defined, but does include a key feature that language is polysystematic (Anderson, 1985), i.e., language is composed of "interacting systems rather than one single system" (Hawkins & Smith, 2001, p. 116). Polysp uses a structural hierarchy to explain phonetic variation. For

[19]Hawkins herself says that "Polysp is not yet specific enough to be a model. It is more a conceptual approach whose aim is to show the value of including fine phonetic detail into any model of how human brains understand speech" (2003, p. 388).

example, anticipatory lip rounding can be present on a higher node in the hierarchy in order to encompass the full duration of lip rounding.

Recall the example of *mistimes* vs. *mistakes* as discussed in section 1.2; the *mis-* of *mistimes* has a "heavier beat" than the *mis-* of *mistakes*. Figure 1.4 shows an example of the differences in acoustic realizations of *mis-* in a Polysp framework, where the differences in morpheme structure give rise to differences in acoustics.

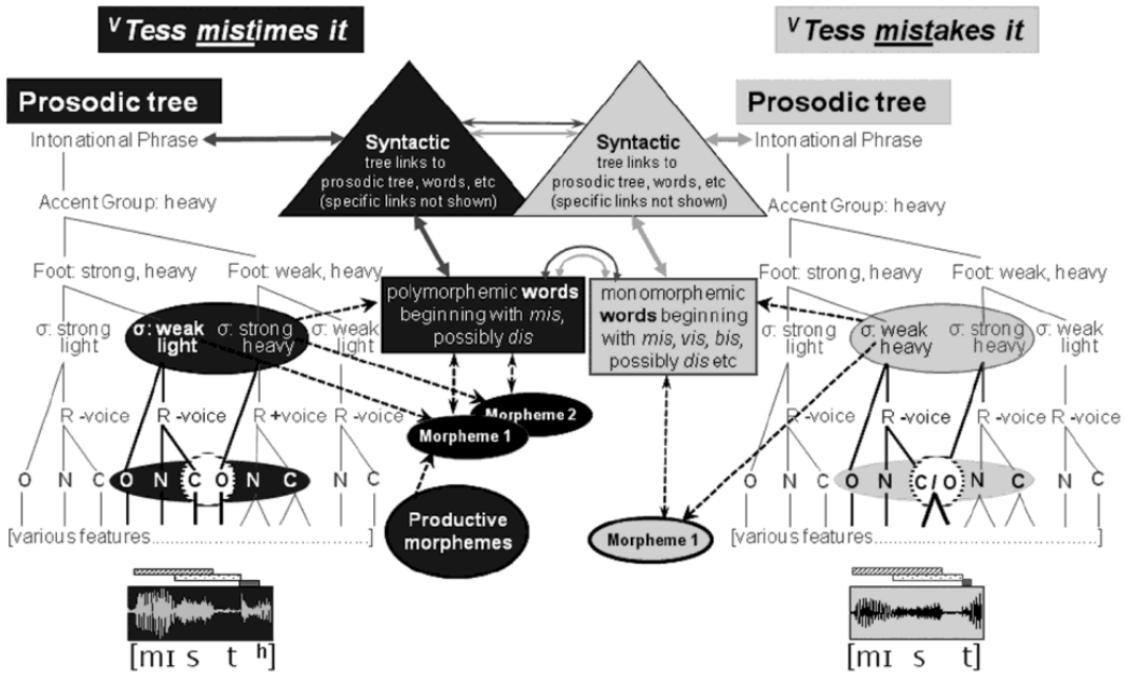

Figure 2. Partial structures (following Ogden *et al.* 2000), describing differences between *mistimes* (left) and *mistakes* (right) in the utterances *Tess mistakes/mistimes it*, with nuclear stress on *Tess*. Bars above the waveforms as in Table 2. Node labels and prosodic tree branches that are the same for both utterances are shown in light grey. Differences are highlighted and in darker shades: the prosodic tree structures corresponding to the two *mist* portions, and hence whether one or two morphemes are identified. Dashed arrows show the parts of the prosodic structure crucial to informing about the particular morphemic structure: two morphemes for *mistimes*, one for *mistakes*. Solid arrows show links to other sources of knowledge.

Figure 1.4: Differences in *mis-*, reproduced from Hawkins (2010, p. 488)

In Figure 1.4, we can see that the differences in the prosodic tree (that the morpheme mis- has a light syllable in *mistimes* and a heavy syllable in *mistakes*) is comes from the differences in polymorphemic words and monomorphemic words.

1.2.6 Concluding Remarks on Phonetic Detail

This section on Phonetic Detail has consisted of a detailed overview of fine phonetic detail followed by an exploration of why this phonetic detail is important to the phonologist. Firstly, this dissertation examines consonant variation (subsection 1.2.1) and some of the ways in which consonants can vary: through coarticulation (subsection 1.2.2), sociolinguistically (subsection 1.2.3), and through reduction (subsection 1.2.4). Then, the entirety of subsection 1.2.5 reviewed the importance of phonetic detail to questions integral to phonological inquiry, followed by an overview of what Abstractionist Theories and Exemplar Theories propose about speech understanding and how they cope with fine phonetic detail. Finally, Hybrid Models are introduced as a solution that is able to preserve the ideas from each theory for which there is the most evidence. Having introduced the topics present in this dissertation, related research questions are given in section 1.3.

1.3 Research Questions

This section outlines the research questions answered by this dissertation. Given the importance of phonetic detail, what properties affect coarticulation, a process that results in variation in phonetic detail? According to Coleman (2003), voicing is one feature that influences coarticulation; however, Davidson (2016) categorizes different patterns of voicing (truly voiceless, bleed, trough, negative VOT, hump, truly voiced), rather than the binary feature [voice]. While Coleman's (2003) study looks at voicing in stops in a *phonological* way, the present study examines the actual *acoustics* of the stop. The research question answered by

Chapter 2 is: do these phonetic voicing patterns affect long-distance coarticulatory properties of the rest of the utterance?

In addition to coarticulation, another aspect of phonetic detail is variation beneath the level of the phoneme. One way in which a stop consonant can vary is in voice onset time (VOT), the length of time between the release of the stop and the onset of modal voicing. While VOT is well-studied, it is has not been studied in the Digital Archive of Southern Speech (DASS) corpus (Kretzschmar et al., 2013; Kretzschmar et al., 2019) and the prior literature on sociolinguistic variables' effect on VOT is contradictory and limited. Therefore, the research question answered by Chapter 3 is: how do both linguistic and sociolinguistic variables (ethnicity, sex, and age) affect the realization of VOT in the DASS corpus?

One application of phonetic detail research is speech technology. Forced alignment is a tool of use to linguists and an offshoot of automatic speech recognition. Automatic speech recognition is typically judged by word error rate, though the individual phone boundaries are of great importance to linguists. So, the question remains: can phonetic detail can be incorporated into forced alignment in order to produce more accurate phone boundaries for linguists? Specifically, Chapter 4 answers the question, will the Montreal Forced Aligner (McAuliffe et al., 2017) be better able to detect an example of subphonemic variation (word-final t/d deletion) if its dictionary is modified? Furthermore, what are the subphonemic acoustic correlates of the stops /t d/ that govern MFA's decision?

This section has summarized the research questions and interests of this dissertation, and the section that follows will provide an overview of the remaining chapters and how they answer these questions.

1.4 Chapter Map

In Chapter 2, the long-distance coarticulatory properties of voicing are investigated. The inspiration to investigate voicing, especially in stops, draws from two sources. First, Coleman

(2003) studies the phonetic detail of American English stops, and the phonological feature [voice] was found to have a long-distance coarticulatory effect. However, voicing is not always realized as categorically present or absent. Davidson (2016) classifies partially voiced stops as having different patterns of voicing: fully voiced, fully voiceless, bleed, trough, negative VOT, and hump. The present study answers the question: How do subphonemic voicing patterns in stops affect the long-distance coarticulatory properties of the stop? Following the methodology of Coleman (2003), LPC coefficients, voicing, F0, harmonics-to-noise ratio, and z-score normalized RMS amplitude were measured in the read data from the Rainbow Passage in the Nationwide Speech Project (Clopper & Pisoni, 2006) in order to examine the stops and their coarticulatory effects. The stops [p, t, k, b, d, g] were classified as one of six voicing patterns, and their surrounding acoustics were measured in 10 millisecond frames that were aligned in time using dynamic time warping. Machine learning classification was then used to classify the voicing pattern based on the surrounding acoustics. Based on its success, Chapter 2 concludes that subtle differences in surrounding acoustics correspond to the voicing patterns of the stop, and therefore, the voicing patterns can affect surrounding segments.

In Chapter 3, the sociophonetic variation of phonetic detail in consonants is examined. One subphonemic property of stop consonants is voice onset time (VOT), the time from the stop closure until modal voicing begins. This study uses the Digital Archive of Southern Speech (Kretzschmar et al., 2013; Kretzschmar et al., 2019), a collection of 64 sociolinguistic interviews recorded between 1968 and 1983, to investigate VOT in Southern speech. AutoVOT (Keshet, Sonderegger, & Knowles, 2014) was used to measure VOT of 206,547 stops (109,023 voiceless stops and 97,524 voiced stops). For voiceless stops, linguistic variables expected to affect VOT of voiceless stops do so, consisting of vowel height, stress, place of articulation, and preceding segment. However, for voiced stops, different linguistics variables have an effect.

In Chapter 4, the Montreal Forced Aligner (McAuliffe et al., 2017) is improved in order to more accurately align tokens featuring word-final t/d deletion. In this study, the dictionary for MFA was modified to include pronunciation variants featuring word-final t/d deletion in order to determine whether MFA would be able to accurately (in comparison with a human transcriber) select the pronunciation variant. For this study, the Buckeye Corpus was aligned, and 23,522 tokens were examined. The forced alignment results agreed with the human transcriber results 71% of the time, which is a rate that is comparable to human intertranscriber agreement. Further, I selected 10% of the dataset to examine manually, marking the subphonemic properties of each /t/ or /d/ in terms of closure, burst (including an epenthetic burst), delayed release, flapping, and glottalization and found the presence of a burst to influence MFA's selection of a /d/, and the presence of a closure, burst, delayed release, flap, or glottalized stop to influence MFA's determination of a /t/'s presence.

Finally, Chapter 5 discusses overall implications of this research and provides general concluding remarks.

Chapter 2

Long-distance Coarticulatory Properties of English Stops

2.1 Introduction

Coarticulation is a well-studied phenomenon in which a segment has some kind of acoustic effect on surrounding segments. Coarticulation can affect segments directly next to it, or affect long-distance segments. Coleman (2003) found voicing to have a long-distance coarticulatory effect on the surrounding segments of the sentence in a study of minimal pairs of sentences differing only in one instance of the phonological feature [voice]. However, voicing is not always realized as categorically present or absent. Davidson (2016) classifies partially voiced stops as having different patterns of voicing: bleed, trough, negative VOT (a voicing pattern that the present study renames "final-third voicing"), and hump. The present study answers the question: How do voicing patterns in stops affect the potential coarticulatory properties of the stop? Following the methodology of Coleman (2003), LPC coefficients, voicing, F0, harmonics-to-noise ratio, and z-score normalized RMS amplitude are measured in the read data from the Rainbow Passage in the Nationwide Speech Project (Clopper & Pisoni, 2006). These variables are measured in the acoustics surrounding the stops present in the sentence,

in order to examine the long-distance coarticulatory effect of these stops. Results show that classification of type of voicing, using long-distance acoustics as features in the classifier, performs better than chance, indicating the existence of meaningful variation. These results have implications for incorporating fine phonetic detail into perceptual models of speech in addition to speech technologies such as forced alignment and automatic speech recognition.

The remainder of this introductory section consists of background literature dealing with coarticulation, how it is measured (subsection 2.1.1), how it can be long-distance (subsection 2.1.2, and how long-distance coarticulation can be measured (subsection 2.1.3); voicing, the different patterns present in obstruents and what influences the realization of these patterns (subsection 2.1.4); how voicing and coarticulation interact (subsection 2.1.5); and concludes with the research questions generated from this literature and answered by this study (subsection 2.1.6).

2.1.1 Measuring Coarticulation

Coarticulation is defined as "changes in articulation and in the acoustic signal induced by one phonetic segment (the trigger) during another one (the target) due to overlap between their articulatory gestures" (Recasens, 2018, n. p.). Many researchers study the articulatory gestures using electropalatographic technology (e.g., Butcher and Weiher (1976), Tabain (2019)), ultrasound (e.g., Irfana and Sreedevi (2019)), and airflow masks (e.g., Coetzee et al. (2019)), but coarticulation can also be studied *acoustically*.

The physical characteristics (i.e., acoustic properties) of the sound wave are what is important to the transmission of a message. The end goal of spoken language is for someone to receive this message: "We speak in order to be heard and need to be heard in order to be understood" (Jakobson & Waugh, 2002, p. 98), and we transmit these messages in an acoustic signal. As Coleman notes, "Articulatory movements are merely a means to an end. In many ways, it doesn't matter very much how the sounds are generated: whether by a

human vocal tract, a speech synthesis program on a computer, a sound recording on CD, or on the TV, or a parrot" (Coleman, 2017, n. p.); the properties of the signal are what is important to the relay of the message, rather than the articulatory movements.

An additional reason to study the acoustic characteristics rather than articulatory gestures is that the same acoustic characteristics can be produced by different articulatory gestures or vocal tract structures. In quantal theory (Stevens, 1972), there are "regions of stability", where speakers' glottal widths can vary within each category of production of a voiceless sound, voiced sound, or glottal stop; different articulatory positions can produce the same voicing quality (Johnson, 2012). Additionally, there is evidence that what auditory qualities speech has are, to some extent, speaker specific due to anatomical differences. For example, Solé found that "some speakers utilize such gestures in a way that allows them to initiate vocal fold vibration during stop closure [...], while similar gestures are executed by speakers without achieving vocal fold vibration" (2018, p. 237), i.e., speakers used different articulatory movements to produce voicing.

To summarize, articulation can be somewhat independent from the acoustic properties of a signal, but the *signal* is what is important in communication. This justifies the approach taken here, to study the acoustics of coarticulation rather than the articulatory gestures.

2.1.2 Long-distance Coarticulation

Coarticulation, as defined in the previous section, affects a previous or subsequent sound segment, but there is evidence that coarticulation can extend beyond adjacent sounds. Öhman provides spectrographic evidence that the vowels in VCV utterances exhibit long-distance coarticulation: "VCV articulations are represented by a basic diphthongal gesture with an independent stop-consonant gesture superimposed on its transitional portion" (1966, p. 151). Furthermore, the phonemes /ɹ/ and /l/ have been shown to have long-distance coarticulatory effects (Tunley, 1999). Kelly and Local compared the phrase "it's a pirate" with "it's a pilot"

and found that "the first utterance is [...] characterized by overall backer resonance. In terms of local resonance the consonantal portions are central or dark in the first, as opposed to half-clear and clear in the second" (1989, p. 74). These long-distance effects of liquids can even be seen up to 5 syllables prior (Heid & Hawkins, 2000).

In the aforementioned study of acoustic correlates of phonological contrasts, Coleman found that words that differed in the phonological feature [voice][1], especially word-finally, had "distributed non-local coarticulatory effects" (Coleman, 2003). Words differing only in the [voice] feature (e.g., mouth_{N} and mouth_{V}) showed long-distance differences in acoustic energies.

2.1.3 Measurement of Long-distance Coarticulation

Long-distance coarticulatory processes of a segment, by definition, are not considered to affect directly adjacent segments. Instead, the coarticulatory influence of one segment affects the phonetic detail of segments further away that the directly adjacent segment. The question is: how far away do we expect coarticulation to occur?

Most long-distance coarticulatory studies of consonants involve liquids. For example, West (1999) found liquids had a long distance coarticulatory effect up to two syllables before the liquid, and Kochetov and Neufeld (2013) found effects of liquids up to five syllables away from the target syllable. Additionally, they found coarticulatory effects of a coronal, [d], but the effects were less pronounced than in the liquids. In Coleman (2003), phones were studied in the context of [t ə C V C ə C] (where the final C is either the [g] of *again* or the [t] in *today* in either *Can you utter/Have you uttered* ____ *again please/today please?*, making the number of phones dependent on whether the contrast was word-initial (two phones to

[1]Traditionally, this feature has been considered [voice] but it has been argued that the feature for the contrast in word-initial stops is [spread glottis], and not [voice] (Beckman, Jessen, & Ringen, 2013; Honeybone, 2005). This point will be returned to in Chapter 3.

the left and four to the right), word-medial (three phones to the left and three to the right) or word-final (four phones to the left and two to the right).

2.1.4 Voicing in English Obstruents

The realization of voicing and factors that affect it are discussed here. It is well-known in phonological theory that voicing contrasts in English are not always recognized by true voicing (i.e., vibration of the vocal folds), and there is a great deal of variation in how voicing is realized, both in terms of vibration of vocal folds and changes in the signal that give the percept of a phonemically voiced segment. The phonetic realization of voicing depends on lexical specification but also on other linguistic factors, such as word position, adjacency to other sounds, lexical stress, and phrase position.

2.1.4.1 Linguistic Factors Affecting Patterns of Voicing

The sections below discuss the following linguistic factors: phone position in the word, position of the phone in the phrase, the lexical stress of the phone, and the proximity of the phone to other sounds.

2.1.4.1.1 Phone Position Phonemically voiced stops word-medially are likely to be voiced (Ladefoged & Johnson, 2014). Accounts of word-initial post-pausal stops are less clear. Typically, these are phonetically voiceless, and the distinction between phonemically voiced sounds and voiceless sounds is due to aspiration (which would then have a longer VOT). In studies, however, some were produced with negative VOT, while others were produced with short lag VOT (Keating, 1984; Lisker & Abramson, 1967). Word-final stops are distinguished by the length of the preceding vowel (Raphael, 1972) and are voiceless.

2.1.4.1.2 Phrase Position Davidson (2016) found that 25% of phrase-initial stops had prevoicing, while about half of stops had voicing phrase-finally. This is explained articulatorily:

for voicing to occur, subglottal pressure must be higher than oral pressure. Phrase-initially, subglottal and oral pressure are similar, and voicing is less likely to occur. At the end of a phrase, the preceding segment is likely to be a sonorant, and it requires less articulatory effort to maintain voicing (Solé, 2018).

2.1.4.1.3 Lexical Stress In Davidson's (2016) study, the stress of vowels preceding and following the obstruent was coded as stressed or unstressed according to the CMU dictionary (Weide, 1998). When preceding vowels were stressed, the obstruents were more likely to be fully voiced. When subsequent vowels were stressed, obstruents were more likely to be partially voiced or devoiced.

2.1.4.1.4 Adjacency to Other Sounds Phonemically voiced stops that are adjacent to a voiced sound typically display phonation, while those that are after a voiceless sound typically do not (Ladefoged & Johnson, 2014). When a stop (or fricative) was preceded by a vowel or approximant, they were only devoiced 11–13% of the time (Davidson, 2016). Adjacency to an obstruent made a voiced stop or fricative less likely to feature phonation (Davidson, 2016; Gonet & Święciński, 2012). Additionally, a preceding nasal made a stop more likely to be voiced, but made a fricative less likely to be voiced (Davidson, 2016).

2.1.4.2 Patterns of Voicing

Voicing is often gradient, and when a sound is partially voiced, there are four patterns to the voicing: 1. Bleed (top left in Figure 2.1), where voicing "bleeds" from the preceding sonorant into the beginning of the consonant, and ceases before the release; 2. Trough (top right in Figure 2.1), where voicing continues from the preceding sonorant, ceases, and then starts again before the release; 3. Negative VOT (bottom left in Figure 2.1), where voicing starts in the middle of the closure, and 4. Hump (bottom right in Figure 2.1), where voicing increases

from the first to second interval but decreases from the second to third interval (Davidson, 2016).

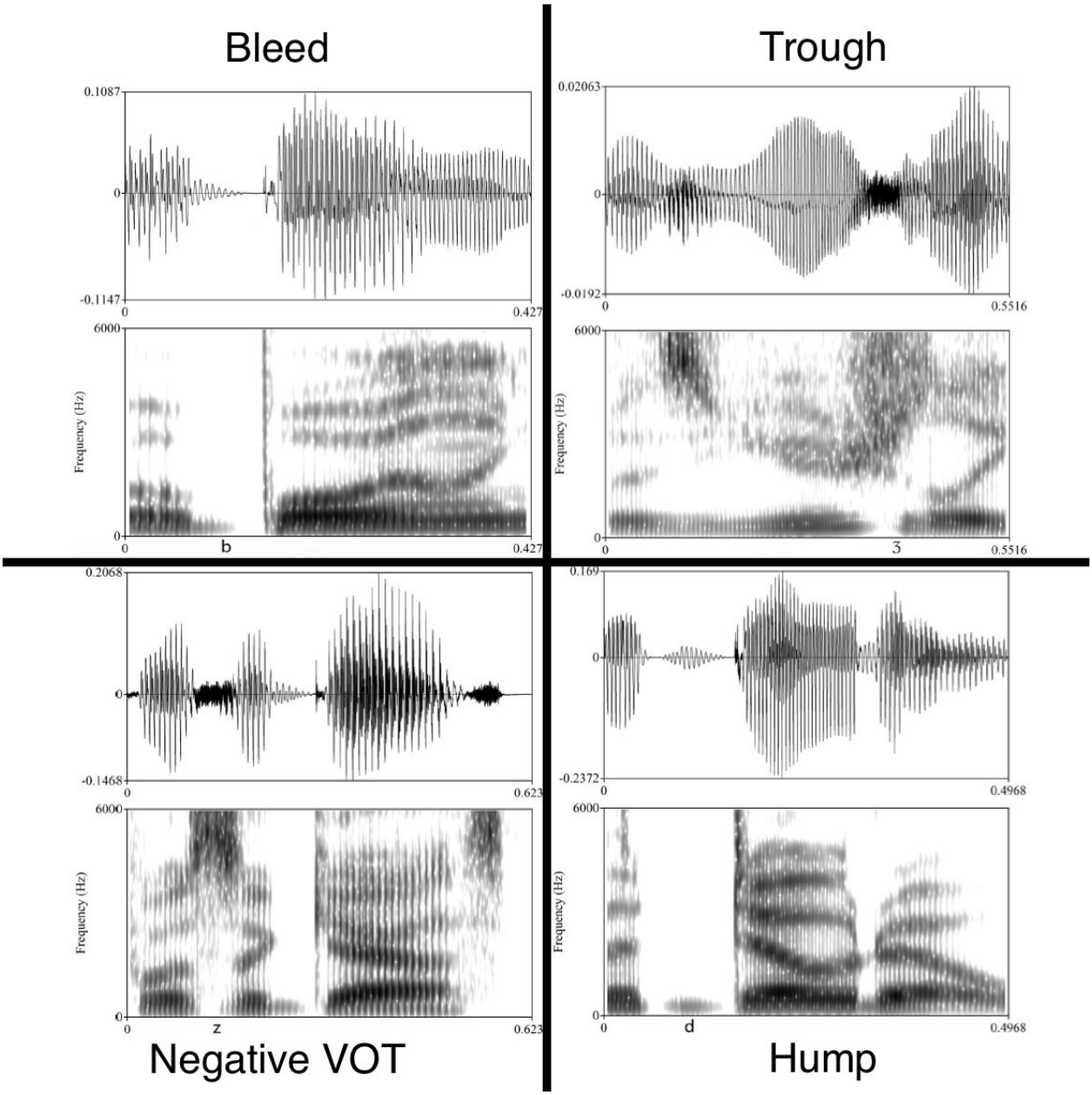

Figure 2.1: Examples of bleed (top left), trough (top right), negative VOT (bottom left) and hump (bottom right), waveforms and spectrograms reproduced from Davidson (2016, p. 43)

The term "negative VOT", as used by Davidson (2016) is different than a canonical definition of negative VOT. Since Figure 2.1 is reproduced from Davidson (2016), the term

"negative VOT" is used here, but the present study will refer to this pattern, where voicing is found in the final third of the stop, as "final-third voicing".

2.1.5 Voicing and Coarticulation

Acoustic analysis is appropriate for the study of the realization of voicing and its relationship to coarticulation: "Acoustic analysis provides information about spectral coarticulatory characteristics [...] associated mostly with tongue movement, nasality, and **voicing** [emphasis mine]" (Recasens, 2018, n. p.). Coleman shows that a contrast in consonant voicing can have a long-distance coarticulatory effect: "The spectrum of /ə/ has more energy before mouth_V than mouth_N, especially above 4 kHz, indicating that /ə/ is breathier before mouth_N than mouth_V, perhaps in anticipation of the voicelessness of the final /ð/" (2003, p. 361), and similar spectral differences were found in sheathe/sheath. Coleman (2003) also found long-distance spectral differences in pairs like *pub/pup*, *robe/rope*, *learned/learnt*, *newt/nude*, *slide/sleight*, *smelled/smelt*, *liege/leech*.

In order to study these acoustic variations, Coleman (2003) recorded one speaker of Southern British English speaking sets of sentences that are minimal pairs, differing only in the phonological feature [voice], e.g., *Can you utter pub again please?* and *Can you utter pup again please?*. In order to represent the acoustic signal, 20 acoustic parameters were extracted in 10 ms frames with 5 ms overlap: 15 LPC coefficients, F_0 and amplitude of voicing (in order to encode voicing), and tilt and root-mean-square (RMS) amplitude (in order to encode noise excitation). These 20 parameters offer a near-complete encoding of the speech signal. In order to compare the contrast between *bed* and *bet*, for example, one could not directly compare frames in time, as one would expect the vowel of *bed* to be longer than *bet*, and so Coleman (2003) used dynamic time warping in order to align these frames in time and "compare apples to apples" (2003, p. 357). Then, the acoustic parameters of each frame of the sentence containing the voiced half of the minimal pair were compared to the

sentence containing the voiceless sound, i.e., the acoustics of *Can you utter pub again please?* were compared to the acoustics of *Can you utter pup again please?*. Coleman considered the difference in the frames to be significant if the confidence intervals of the two groups were disjoint "where the lower confidence limit of one word's tokens is greater than the upper confidence limit of the other word's tokens" (2003, p. 359).

2.1.6 Research Questions

Given that there are different patterns of voicing, and voicing has a long-distance coarticulatory effect, my research question is as follows: when stops have different patterns of voicing (completely voiceless, bleed, trough, final-third voicing, hump, completely voiced), are there acoustic differences in the rest of the utterance? The answer to this question builds on Davidson (2016, 2018) by determining the coarticulatory consequences of the voicing patterns she defines, and builds on Coleman's (2003) work by expanding the categories of voicing to be phonetically, rather than phonologically, based. The null hypothesis in this study is that the acoustics in the utterance are the same regardless of voicing pattern, and the alternative hypothesis is that the acoustics in the utterance are different based on voicing pattern. This hypothesis is tested following Coleman (2003) via acoustic analysis of corpus data which are then aligned in time, though this methodology differs in a two crucial ways: 1. data was collected from 60 speakers, instead of one, and 2. machine learning classifiers were used to classify acoustic data according to voicing pattern, the argument being that if the algorithm can classify the type of voicing as a result of the long-distance acoustics, then there is potentially long-distance coarticulation based on the voicing pattern. Additionally, in order to use a consistent measurement, the present study measures two syllables in each direction from the syllable containing the target consonant based on West's (1999) measure of consonants.

2.2 Methodology

This methodology section first describes the data used from the Nationwide Speech Project corpus (Clopper & Pisoni, 2006) (subsection 2.2.1). It also discusses the variables in the study: how voicing patterns of the stops in question are determined in subsection 2.2.2 and the acoustic and linguistic variables used in the study in subsection 2.2.3. This section also describes the dynamic time warping process (subsection 2.2.4), which needed to be done so that frames containing the acoustic measurements are aligned in time, in addition to providing a detailed account of what parts of the sentences in the corpus were examined (subsection 2.2.5). Finally, the process of classification (i.e., prediction of the independent variable using dependent variables as features) is outlined in subsection 2.2.6.

2.2.1 Nationwide Speech Project Corpus

The Nationwide Speech Project (Clopper & Pisoni, 2006) contains data from 60 talkers. These talkers included ten from each of the six dialect regions as designated by Labov, Ash, and Boberg (2006): New England, Mid-Atlantic, North, Midland, South, and West. The talkers are also balanced with respect to gender. The talkers are young, ranging from ages 18 to 25; speak English; and are white. The materials used for this study are the read Rainbow Passage (Fairbanks, 1940). First, the audio .aiff files were converted to wav using SoX (Bagwell & Norskog, 2015). The files were force aligned using WebMaus services (Kisler, Reichel, & Schiel, 2017) in order to produce Praat (Boersma & Weenink, 2018) TextGrids with boundaries for individual phones and words. Forced alignments were confirmed accurate at the sentence level for all speakers, and these files were separated into sentences, based on orthography (i.e., by where periods are), but also by visually inspecting the audio files to confirm that sentence boundaries are where natural pauses occur. These sentences are listed in Appendix A. According to the README file of the Nationwide Speech Project, "Most

of the recordings of the Rainbow and Goldilocks passages include some disk-skipping at the beginning", and so these beginning sentences were excluded. The Nationwide Speech Project was used in order to have different speakers reading the same passage, so that acoustics of the same sentences can be compared.

2.2.2 Measuring Voicing of Stops

In order to answer the research question of variation due to *voicing* in stops, the following section details how the measurement of voicing of the stop was obtained. The voicing of each stop in the corpus was calculated with a Praat script using the Praat Voice Report (Boersma & Weenink, 2018). In order to determine the proportion of unvoiced frames over a defined period of time, the Voice Report feature requires a Pitch Object and a PointProcess Object. Following Eager (2015), Pitch Objects and PointProcess Objects were created using sex-specific values. The pitch floor was defined as 70 Hz for males and 100 Hz for females, and the pitch ceiling was defined as 250 Hz for males and 300 Hz for females. The time step was defined as 0.001s. Other parameters (maximum number of candidates, silence threshold, voicing threshold, octave cost, octave-jump cost, voiced/unvoiced cost) were left at their standard values defined in the Praat manual (Boersma & Weenink, 2018). From these Pitch and PointProcess Objects, Praat calculated the proportion of unvoiced frames in the first, second, and third interval of the stop in order to get a measure of how the voicing changes over the duration of the stop, allowing us to assign a pattern of voicing.

Following Davidson (2016), both a categorical measure and the shape of voicing were obtained. Tokens were classified categorically as "voiced" (>90% identified as voiced), "unvoiced" (<10% identified as voiced), or "partially voiced" (10-90% voiced). To assess the shape of voicing, voicing was measured in three intervals of each sound. Shape of voicing was classified as "bleed" (voicing decreases from the first interval to the third interval), "final-third voicing" (voicing increases from the first interval to the third interval), "trough" (first and

third intervals have greater voicing than the second interval), and "hump" (voicing increases from the first to second interval, and voicing decreases from the second to third interval).

Figure 2.2 shows examples of these four voicing patterns from the present data. In the top left, there is an example of bleed in the word-final [t] of *sunlight* from speaker at0 (a female speaker from the Mid-Atlantic); in the top right, there is an example of bleed in the word-initial [b] of *boiling* from speaker at3 (a male speaker from the Mid-Atlantic); in the bottom left, in the word-initial [p] of *pot* from speaker so6 (a female speaker from the South), there is an example of final-third voicing. Finally, in the bottom right, there is an example of hump, as there is voicing present in the aspiration, in the word-initial [k] of *colors* from speaker no9 (a female speaker from the North).

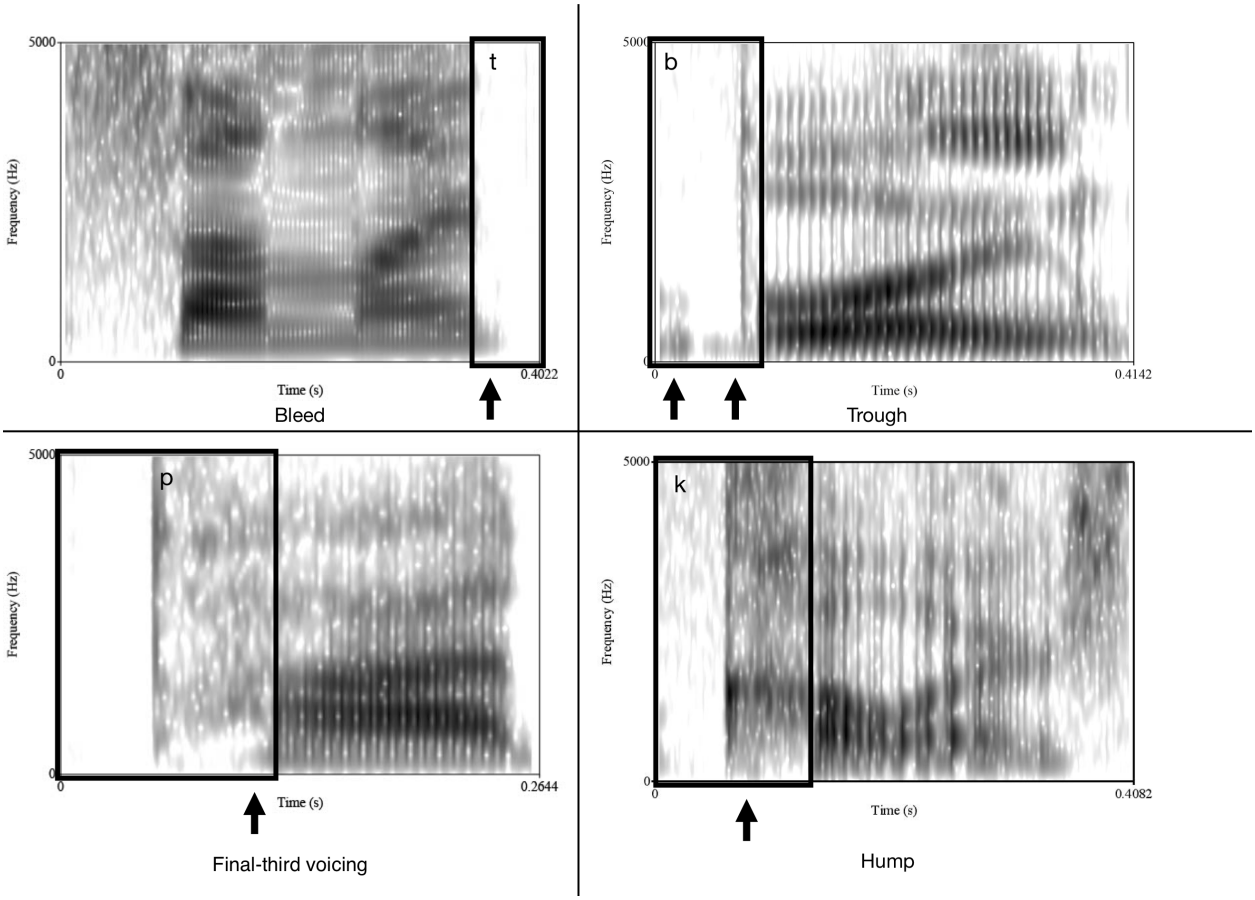

Figure 2.2: Examples of bleed (top left), trough (top right), final-third voicing (bottom left), and hump (bottom right) from the Nationwide Speech Project (Clopper & Pisoni, 2006)

2.2.3 Variables

The following section details the variables recorded in this study. The variables can be divided into two groups: acoustic variables and linguistic variables. The acoustic variables encompass information about the acoustic signal that surrounds the stop in question, and they include LPC coefficients, voicing, pitch, RMS amplitude, and harmonics-to-noise ratio. The linguistic variables the position of the stop in the word, the position of the stop in the phrase, the lexical stress associated with the word containing the phone, and the stops' following and preceding segments.

2.2.3.1 Acoustic Variables

After obtaining the voicing of each stop, the next step was to get the acoustics surrounding these stops. For each 10 ms frame (with 5 ms overlap), the following 19 acoustic parameters were measured: 15 LPC coefficients, proportion of voicing, average F0, average harmonic-to-noise ratio, and z-score normalized RMS amplitude.

2.2.3.1.1 LPC Coefficients To obtain the LPC coefficients, Praat (Boersma & Weenink, 2018) was used to resample the .wav files to 16,000 Hz, and a 15 order LPC model was obtained, yielding 15 LPC coefficients for each 10 ms frame of time (with 5 ms overlap).

2.2.3.1.2 Voicing and Pitch For voicing, the Praat Voice Report feature was used once again to get proportion of unvoiced frames of the entire utterance, and the settings to get voicing and pitch are described in subsection 2.2.2. The proportion of unvoiced frames was measured for 10 ms frames (with 5 ms overlap) and converted to a measure of voiced frames by taking the inverse. For pitch, an average over the 10 ms frame was taken.

2.2.3.1.3 RMS Amplitude The RMS of each frame was taken, and this measurement was normalized by subtracting it from the average RMS for the entire .wav file, which was then divided by the standard deviation.

2.2.3.1.4 Tilt/Harmonics-to-noise Ratio Coleman (2003) uses spectral tilt, measured by finding the first and second harmonics, to measure breathiness; however, Simpson (2012) found that there is influence from vowel nasality on these harmonics. The present study takes the recommendation of Styler (2013) who suggests using harmonics-to-noise ratio to measure breathiness instead of spectral tilt based on Simpson (2012).

2.2.3.2 Linguistic Variables

In addition to the acoustic variables, other linguistic variables Davidson (2016) found to have an effect on voicing, word position, phrase position, lexical stress, and adjacency to other sounds, were also recorded.

2.2.3.2.1 Word Position Word position refers to the position of the phone within the word, and was classified as either word-initial, word-medial, or word-final.

2.2.3.2.2 Phrase Position Phrase position refers to the position of the phone within the phrase (in this case, taken to be a clause), and was classified as either phrase-initial, phrase-medial, or phrase-final. All but two phones were phrase-medial, and so this variable was not included in subsequent analysis[2].

2.2.3.2.3 Lexical Stress The lexical stress of the words was determined by the CMU dictionary (Weide, 1998). Since the CMU dictionary assigns stress to a vowel, pronunciations were syllabified in order to assign stress to the syllable containing the vowel. These

[2]As this variable was only included in the final classifier model, it is unlikely to affect results.

pronunciations were syllabified according to the Maximal Onset Principle, legal sequences in English, and morphological factors[3].

2.2.3.2.4 Adjacency to Other Sounds The preceding and following sounds were classified as either a stop, fricative, affricate, palatal, nasal, approximant, vowel, or a silence, as determined by the forced alignment.

2.2.4 Dynamic Time Warping

Once these 20 acoustic variables were obtained, dynamic time warping was used to time-align them. Dynamic time warping is an algorithm that finds an optimal alignment between two time series, and one is mapped to another. Since the speaking rates are different, dynamic time warping was used so that all the sounds have an equal number of frames in time; so that we can compare "apples to apples" (Coleman, 2003, p. 357).

In order to perform the dynamic time warping, matrices of each sentence were used. Since there are 60 instances of each sentence, there are 60 matrices, where each matrix is one sentence, the rows are 10 ms frames, and the columns are the variables (LPC coefficients, etc.). The normalized distance between each pair of matrices was found using only the LPC coefficients[4], and these distances were added together to get the total normalized distance. This total normalized distance was used as the reference that each of the other matrices was warped to. Using the R package 'dtw' (Giorgino, 2009), the optimal warping curve was calculated between each of the 59 sentences and the least dissimilar sentence from all others. This warping curve was used to subscript each matrix, with the result being that all matrices of the same sentence have the same number of rows (i.e., the same number of frames in time).

[3]For example, in the word *plating* (*plat* + *ing*), "morphological considerations may pull the consonant to the first syllable" (Eddington, Treiman, & Elzinga, 2013, p. 57).

[4]Only LPC coefficients were used because the other variables (proportion of voiceless frames, F0, harmonics to noise ratio, z-score RMS) are not well-aligned in time.

When time series data is multidimensional, there are two ways to compute the warping path, DTW_I ("I" referring to "independent") and DTW_D ("D" referring to "dependent"). In independent DTW, the distance is independently computed for each dimension (i.e., multiple warping paths): $DTW_I(Q,C) = DTW(Q_xC_x)+DTW(Q_yC_y)$, and in dependent DTW, there is only a single warping path computed: $DTW_D(Q,C) = DTW(Q_xQ_y, C_x, C_y)$. Dependent DTW was used in this case, as dependent DTW is best for when a "physical process affects the time series simultaneously" (Mueen & Keogh, 2016, slide 59). Typically, data must be normalized when DTW is used for classification, but there is evidence that you can combine raw and normalized data (Luczak, 2018), and in this case, we can just compare raw audio to raw audio instead of using the normalized data for dynamic time warping *classification*.

In the measurements of voicing in the data, Praat (Boersma & Weenink, 2018) often reports that the pitch and harmonics-to-noise ratio of voiceless sounds are "—undefined—". Initially, these values were treated as "NA" instead of having a value of zero. However, dynamic time warping cannot handle data with NAs. Imputation (i.e., replacing missing data) can handle data with NAs in several different ways. There is listwise deletion, where missing data is deleted, but this method does not seem advisable in time series data where there are values sampled evenly in time. There are also imputation methods that have to do with data insertion, but again, we know that the values for pitch and harmonics-to-noise ratio do not exist, and selecting data with values (such as mean substitution or hot/cold deck, where information is gathered from the same or different data sets) for these variables is not advisable either. The imputation of multivariate time series depends on the data being missing completely at random, and this missing data is not random; it is missing because there is no pitch in voiceless sounds. Because no form of imputation works with this data, and because NAs cannot be handled by dynamic time warping, these values were changed to zero using the R package 'timeSeries' (Wuertz, Setz, & Chalabi, 2017).

2.2.4.1 Calculation of Frame Number

In order to understand what frame of time corresponds to what acoustics, the frame number was calculated by subtracting the start time of LPC coefficient measurement from the time measured. The start time of LPC coefficient measurement does not always occur at the exact start of the file, because of the nature of taking measurements over evenly-spaced frames. Milliseconds were divided by five in order to get frames, and then one was added for a start frame and one was subtracted to get the end frame for that time. Let x be the time in seconds, and y be the frame number. To get the frame that starts at that time, the following equation was used $y = (x - 0.01) * 1000 \div 5 + 1$. To obtain the frame that ends at that time, the formula is $y = (x - 0.01) * 1000 \div 5 - 1$. Figure 2.3 shows the two syllables on each side of the target syllable, the start and end time of the five-syllable segment in the centroid, and the corresponding frame numbers to these times.

2.2.5 Sentences

Figure 2.3 gives an overview of each of the stops measured and the linguistic variables in each sentence. Only stops that were judged by the forced aligner to be produced by all speakers were included. For example, the [d] of *and* and the [t] of *into* of sentence 2 were excluded for this reason.

Sentence	Phone	2 syllables before	target syllable	2 syllables after	Start time	Preceding phone start	Start phone	End phone	Following phone end	End time	Start frame	Preceding phone start frame	Phone start frame	Phone end frame	Following phone end frame	End frame	Sentence warped to
2	[b]	The rain	bow	is a	0.01	0.26	0.33	0.36	0.61	0.98	1	51	65	69	119	193	ne3
2	[d]	is a[1]	di	vision (division)	0.61	0.77	0.98	1.11	1.15	1.5	121	153	195	219	227	297	ne3
2	[b]	many	beau	tiful (beautiful)	2.45	2.65	2.73	2.83	2.86	3.14	489	529	545	563	569	625	ne3
2	[k]	tiful	co	lers (colors)	2.91	3.11	3.14	3.28	3.31	3.64	581	621	627	653	659	725	ne3
3	[t]	these	take	the shape	0.01	0.1	0.15	0.28	0.37	0.86	1	19	29	53	71	169	so4
3	[t]	arch with	its	path high	1.81	2.93	2.98	3.01	3.04	3.94	361	585	595	599	605	785	so4
3	[p]	with its	path	high a (above)	2.81	3.01	3.04	3.18	3.39	3.98	561	601	607	633	675	793	so4
3	[b]	path a	bove	and its	3.04	3.94	3.98	4.08	4.25	5.06	607	787	795	813	847	1009	so4
3	[t]	bove and	its	two ends	3.98	4.95	4.99	5.03	5.06	5.61	795	989	997	1003	1009	1119	so4
3	[t]	and its	two	ends a (apparently)	4.86	5.03	5.06	5.19	5.33	5.65	971	1005	1011	1035	1063	1127	so4
3	[p]	ends a	ppar	ently	5.33	5.61	5.65	5.81	5.85	6.12	1065	1121	1129	1159	1167	1221	so4
3	[b]	ently	be	yond the	5.9	6.05	6.12	6.2	6.27	6.99	1179	1209	1223	1237	1251	1395	so4
4	[k]	is a	ccor	ding to	0.28	0.45	0.48	0.61	0.65	1.2	55	89	95	119	127	237	ne2
4	[b]	gend a	boil	ing pot	1.86	2.55	2.58	2.69	2.8	4.03	371	509	515	535	557	803	ne2
4	[p]	boiling	pot	of gold	2.58	2.91	3.85	3.9	4.03	4.62	515	581	769	777	803	921	ne2
4	[g]	pot of	gold	at one	3.85	4.08	4.15	4.28	4.51	5.75	769	815	829	853	899	1147	ne2
5	[p]	-	Peo	ple look	0.01	0.01	0.01	0.03	0.1	0.68	1	1	1	3	17	133	we4
5	[p]	peo	ple	look but	0.01	0.03	0.1	0.17	0.2	1.29	1	5	19	31	37	255	we4
5	[k]	people	look	but no	0.01	0.42	0.53	0.68	1.22	1.42	1	83	105	133	241	281	we4
5	[b]	ple look	but	no one	0.1	0.15	1.12	1.22	1.26	1.6	19	29	223	241	249	317	we4
6	[k]	a man	looks	for some	0.09	0.48	0.57	0.63	0.73	1.03	17	95	113	123	143	203	so4
6	[b]	something	be	yond his	0.87	1.14	1.22	1.28	1.32	1.66	173	227	243	253	261	329	so4
6	[k]	he is	look	ing for	3.27	3.66	3.69	3.76	3.79	4.03	653	731	737	749	755	803	so4
6	[p]	for the	pot	of gold	3.89	4.07	4.1	4.26	4.32	4.83	777	813	819	849	861	963	so4
6	[g]	pot of	gold	at the	4.1	4.4	4.45	4.55	4.62	5.03	819	879	889	907	921	1003	so4
6	[b]	the rain	bow	-	5.42	5.63	5.71	5.76	5.92	5.92	1083	1125	1141	1149	1181	1181	so4

Figure 2.3: Stops measured and corresponding warped frames of time

Table 2.1: Centroid Speakers

Sentence	Centroid speaker
2	Male speaker from New England (identified as ne3 in the data)
3	Male speaker from the South (so4)
4	Male speaker from New England (ne2)
5	Male speaker from the West (we4)
6	Male speaker from the South (so4)

The centroid speaker for each sentence (i.e., what speaker spoke the sentence that all other sentences were warped to) is given in Table 2.1.

2.2.6 Classification

The goal of the classifier is to classify the pattern of voicing during the phone (bleed, trough, final-third voicing, hump, voiceless, voiced) based on the acoustics present in each frame. In total, 5,917 frames were used for analysis. Classifying the type of voicing by each frame in the data is appropriate because each frame of time may not be affected by coarticulation. For example, in word-final [ð]/[θ] contrast (*Can you utter mouth again please?*), Coleman found differences in the preceding vowels [aʊ] and [ə] (but not [m][5]). These differences were measured for each parameter in each 5 ms frame. In other words, we might expect the effect of long-distance coarticulation to be inconsistent over all the frames; there are particular frames that get classified better (where long-distance coarticulation might be happening) and frames that get classified worse (the acoustics of these frames do not vary). Additionally, all acoustic data from the phone itself and from the preceding and following phone was analyzed in separate classifiers, in order to avoid classifying the type based on data from the phone, which would obviously be a strong predictor of the type of voicing, and in order to avoid capturing the relationship between non-long distance coarticulation and type of voicing. In

[5] *Utter* is pronounced as [ʌtə], as Coleman (2003) uses British English, a non-rhotic dialect.

the classifier to test for long-distance coarticulation, for example, in the portion of sentence 2, *The rainbow is a*, the acoustics of this entire segment are used except for the frames that correspond to [nboʊ].

In order to organize data, RStudio (R Core Team, 2018; RStudio, 2017) and the R package 'tidyverse' (Wickham, 2017) were used. Then, data were partitioned into 80% training and 20% testing, feature scaled in order to avoid larger values having more of an effect on the classifier. The training data was subjected to the Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) in order to achieve a 1:1 balance of the predicted group in order to balance the data set. The classifier was made using the R package 'RWeka' (Hornik, Buchta, & Zeileis, 2009) an R wrapper for Weka (Frank, Hall, & Pal, 2016), a machine-learning software package, along with the R packages 'rJava' (Urbanek, 2019), 'caTools' (Tuszynski, 2019), 'DMwR' (Torgo, 2010), to make a classifier. Weka's J48 classifier was used, which uses the C4.5 algorithm (Quinian, 1993) for building decision trees. The results of the classifier are evaluated using the Kappa statistic.

2.2.6.1 Why Kappa?

A Kappa statistic measures inter-rater agreement, but it also takes into account random chance, making it a better measure than precision and accuracy alone. To see why the Kappa statistic is necessary, consider the following example. Say that we built a machine learning classifier that classified the phones [b] and [p]. If our "ground truth" is that we have 95 tokens of [p], and 5 tokens of [b], the machine learning algorithm, simply by classifying all the tokens as [p], would achieve an accuracy[6] of 0.95, a precision[7] of 0.95, and a recall[8] of 1. The Kappa statistic takes into account random chance, as its formula in shown in Equation 2.1.

[6] $\frac{\text{correct identification of [p] and [b]}}{\text{all tokens}}$, also $\frac{TP+TN}{TP+TN+FP+FN}$

[7] $\frac{\text{correct identification of [p]}}{\text{all identification of [p]}}$, also $\frac{TP}{TP+FP}$

[8] $\frac{\text{correctly selected instances of [p]}}{\text{collect identification of [p] and [b]}}$, also $\frac{TP}{TP+TN}$

Expected accuracy is calculated as shown in Equation 2.2, and the calculation of marginal frequency is shown in Equation 2.3.

$$\kappa = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}} \tag{2.1}$$

$$\text{Expected accuracy} = \frac{\text{Marginal frequency of one class+marginal frequency of second class}}{\text{Total number of instances}} \tag{2.2}$$

$$\text{Marginal frequency} = \frac{\text{Instances labeled as a certain class by classifier} \cdot \text{Instances labeled as a certain class according to the "ground truth"}}{\text{All instances}} \tag{2.3}$$

Following our example, the marginal frequency for [p] is $\frac{100*95}{100} = 95$, the marginal frequency for [b] is $\frac{0*5}{100} = 0$, which means our expected accuracy is $\frac{95+0}{100} = 0.95$. If we then use this to calculate Kappa, $\frac{0.95-0.95}{1-0.95} = 0$, we get a Kappa value of 0, showing that the classifier did not perform better than chance.

2.3 Results

First, to confirm the validity of this methodology, the phone (b, d, g, p, t, k) was predicted from the LPC coefficients of the frames corresponding to the phone. The data was prepared as described above. When predicting phone in the test data ($n = 5616$), the classifier predicted correctly 87.07% of the time, and a confusion matrix is given in Table 2.2, where rows sum to the actual instances of phones and columns are predicted phones.

Adding the other acoustic variables increased performance to 89.69%, but removing each variable one at a time (kappa statistics of which are shown in Table 2.3) revealed that only

Table 2.2: Confusion matrix: Phone labels

		Classified as					
		b	d	k	p	t	g
Stop	b	**1284**	0	58	48	26	0
	d	29	**221**	0	0	0	0
	k	0	76	**1169**	15	0	0
	p	35	0	14	**1141**	132	46
	t	98	0	0	11	**635**	0
	g	0	0	0	88	0	**440**

F0 and voicing has an impact on the model, so in future models, harmonics-to-noise ratio, and z-score RMS were excluded.

Table 2.3: Kappa statistic values

Variable removed	**Performance**
Voicing	0.8849
F0	0.8872
Harmonics to noise	0.8932
Z-score RMS	0.8941

The confusion matrix for the final model is given in Table 2.4 (including LPC coefficients, frame number, voicing, and F0), which performs at 89.69%, $\kappa = 0.8709$, considered 'almost perfect' by Landis and Koch (1977). To further point out the value of using kappa, the expected accuracy is 0.2, meaning that the observed accuracy is much greater than the expected accuracy, which is reflected in the kappa value.

Figure 2.4 shows the number of tokens for each voicing pattern, separated by stop.

Completely voiced and hump voicing patterns had the fewest tokens, and the voiced pattern primarily features the phones /b d t/. The bleed voicing pattern had the highest count, followed by completely voiceless, trough, and final-third voicing. This figure shows the number of *tokens* in the Nationwide Speech Project corpus (Clopper & Pisoni, 2006), i.e., the number of stops measured using the Praat (Boersma & Weenink, 2018). However,

Table 2.4: Confusion matrix: Phone model, final

		Classified as					
		b	d	k	p	t	g
Stop	b	**1332**	0	58	0	26	0
	d	79	**221**	0	0	0	0
	k	0	76	**1169**	15	0	0
	p	0	0	14	**1174**	134	46
	t	98	0	0	11	**635**	0
	g	0	0	0	22	0	**506**

subsequent results examine *frames* of time and whether or not individual frames of time contained information that machine learning algorithms could use to predict the voicing pattern of the stop.

Next, since some level of coarticulation in the preceding and following segments due to voicing might be expected, we can build a classifier that predicts voicing (voiced, voiceless, or partially) from the each frame's LPC coefficients ($n = 7118$), F_0 and voicing of the preceding and following segments. In other words, for each 10 ms frame of acoustic measurements that directly surround the stop, the classifier predicts the voicing of the stop. This classifier has an accuracy of 68.80%[9], $\kappa = 0.4572$, considered moderate (Landis & Koch, 1977), and a confusion matrix is given in Table 2.5. We would expect the items that are most phonetically similar to be confused, and this is borne out in the results. Unvoiced and voiced sounds are more frequently incorrectly categorized as partially voiced.

Further analysis by type of voicing (expanding the partially voiced category into the four options) in addition to looking at acoustic variables from the two syllables before and after the syllable containing the phone (but omitting the preceding phone, the phone itself, and the following phone), ($n = 53,136$), yields a classifier that performs at 51.19%[10], $\kappa = 0.3897$,

[9]Expected accuracy of 42.6%.
[10]Expected accuracy of 20.02%.

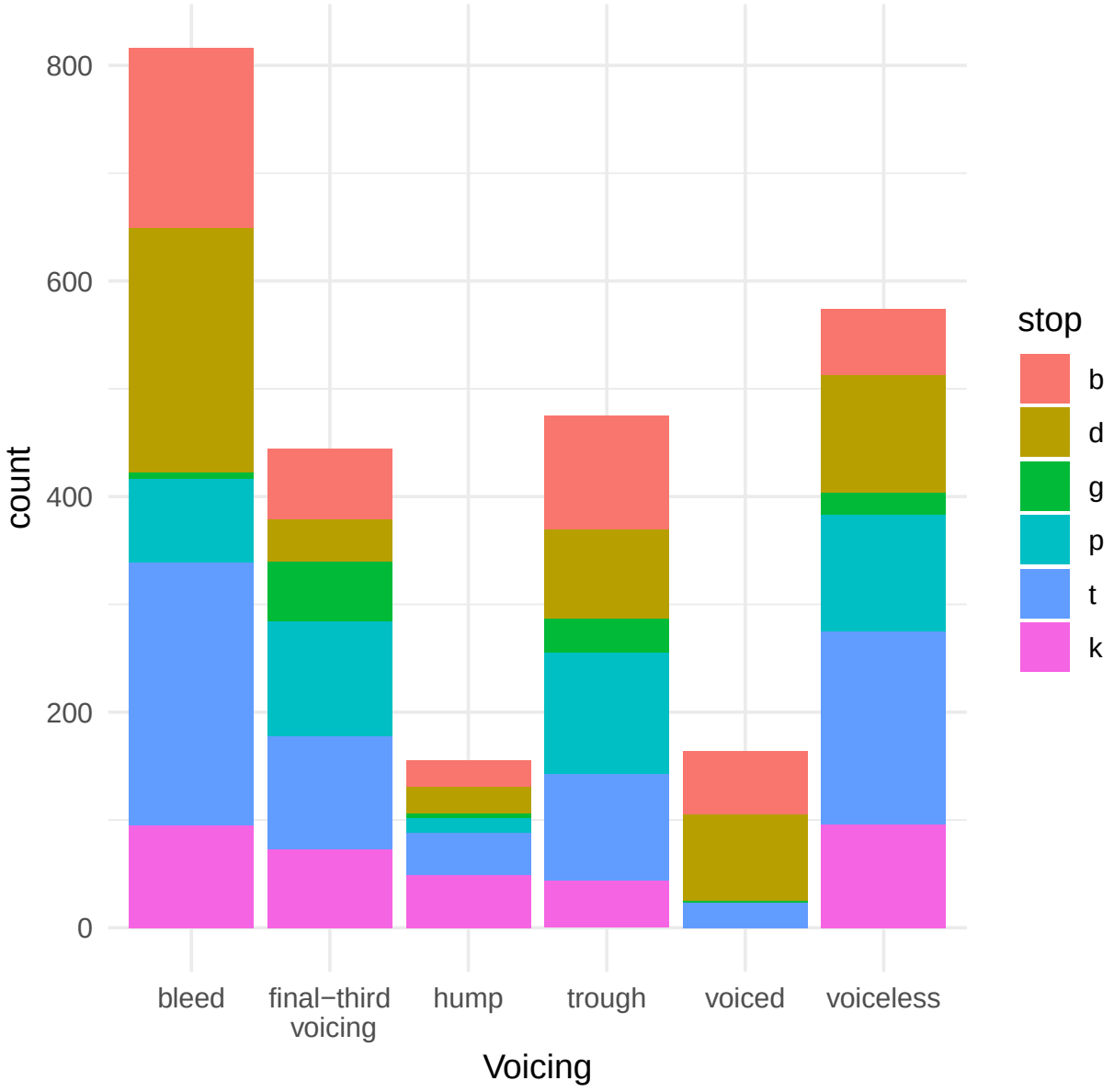

Figure 2.4: Number of tokens exhibiting each voicing pattern

considered the higher end of 'fair' (Landis & Koch, 1977), and a confusion matrix is given in Table 2.6.

Additionally, when the variables type of coarticulation (carryover or anticipatory), preceding sound, following sound, position of phone (word-initial, word-medial, word-final), stress,

Table 2.5: Confusion matrix: Coarticulation in frame-by-frame classifications

		Classified as		
		partially voiced	unvoiced	voiced
Voicing	partially voiced	**2030**	741	290
	unvoiced	878	**2570**	268
	voiced	21	7	**313**

Table 2.6: Confusion matrix: Voicing pattern in frame-by-frame classifications

		Classified as					
		bleed	hump	final-third voicing	trough	voiced	voiceless
Voicing	bleed	**8157**	401	1547	1773	1340	1650
	hump	566	**1139**	370	299	347	322
	final-third voicing	1878	328	**4888**	1337	976	1210
	trough	1721	327	1338	**5993**	1104	1006
	voiced	63	11	22	54	**1648**	68
	voiceless	1894	355	1358	1327	943	**5376**

and region, are added, the classifier performs at 68.49% accuracy[11], $\kappa = 0.6025$, considered the higher end of 'moderate' (Landis & Koch, 1977), and a confusion matrix is given in Table 2.7. Because the data is unbalanced, it is useful to look at the same confusion matrix featuring percents, as shown in Table 2.8. Finally, a summary of classifiers built and their results are given in Table 2.9.

[11]Expected accuracy of 20.02%. This number is slightly different from the expected accuracy of the previous classifier, though it does not show due to rounding.

Table 2.7: Confusion matrix: Voicing pattern in a frame-by-frame classification with linguistic variables

		Classified as					
		bleed	hump	final-third voicing	trough	voiced	voiceless
Voicing	bleed	**10,646**	229	775	1097	748	1373
	hump	311	**1847**	371	194	107	213
	final-third voicing	830	310	**7183**	1013	304	978
	trough	1156	232	1010	**7720**	627	744
	voiced	58	10	13	13	**1771**	1
	voiceless	1459	287	1141	797	342	**7227**

Table 2.8: Confusion matrix: Voicing pattern in a frame-by-frame classification with linguistic variables (percents)

		Classified as					
		bleed	hump	final-third voicing	trough	voiced	voiceless
Voicing	bleed	**71.6%**	1.5%	5.2%	7.4%	5.1%	9.2%
	hump	10.2%	**60.7%**	12.2%	6.3%	3.5%	7.0%
	final-third voicing	7.8%	2.9%	**67.6%**	9.7%	2.9%	9.2%
	trough	10.1%	2.0%	8.8%	**67.2%**	5.5%	6.5%
	voiced	3.1%	0.5%	0.7%	0.7%	**94.9%**	0.0%
	voiceless	13.0%	2.6%	10.1%	7.1%	3.0%	**64.2%**

Table 2.9: Classifier results

Classifier	Description	Accuracy	Kappa Statistic
Classifier 1	Classify phone based on the acoustics during the phone	89.69%	0.8709 (Almost perfect)
Classifier 2	Classify voicing of phone by coarticulation	68.80%	0.4572 (Moderate)
Classifier 3	Classify voicing pattern of phone by long-distance coarticulation	51.19%	0.3897 (Higher end of fair)
Classifier 4	Classify voicing pattern of phone by long-distance coarticulation and linguistic variables	68.49%	0.6025 (Higher end of moderate)

2.4 Discussion

This study has extracted acoustic information from five sentences from 60 speakers, aligned it in time, and used a classifier to predict the voicing pattern (bleed, hump, final-third voicing, trough, voiced, voiceless) of each stop from the acoustics surrounding it. If the null hypothesis, that the acoustics in the utterance are the same regardless of voicing pattern, was true, we would expect a kappa statistic value of 0 (meaning, it has performed at chance) for the classifier that predicts type of voicing from acoustics. Because we do get a kappa statistic value of 0.3897, fair, the classifier is performing above chance, and we can reject the null hypothesis and say that the acoustics of the utterance are in some way affected by the type of voicing of the segment.

Looking back at Table 2.7, the classifier does best on the completely voiced segments. We would expect confusion on phonetically similar results and accuracy on less phonetically similar results. For example, the voicing pattern "hump" is least like "trough"; in fact, they are exact opposites. This prediction is borne out in the classifier: only 194 instances of "hump" are classified as "trough" from a total of 3043 instances of "hump", a rate of 6.4%, and 232 instances of "trough" are classified as "hump" out of 11,489 instances of "trough", a rate of 2%. Similarly, we would expect accuracy in the distinction of "bleed" and "final-third voicing" patterns. There were 775 instances of "bleed" classified as "final-third voicing" from a total of 14,868 actual instances of "bleed", a rate of 5.2%. Additionally, there were 830 instances of "final-third voicing" classified as "bleed" from a total of 10,617 actual instances of "final-third voicing", a rate of 7.8%. These error rates can be compared with the overall error rate of the classifier, which is 31.5%.

These results show that long-distance phonetic detail exists, and these details can potentially be used to recover information about linguistic structure. These results are indicative of what Coleman calls "phonetic influence":

> Phonological contrasts are not in general associated with segment-sized stretches of speech. On the contrary, even ordinary 'phonemic' contrasts are phonetically realized by a combination of short-time and more extended phonetic correlates. Some of the short-time correlates might well be termed 'sub-segmental' (2003, p. 365).

If a listener is able to recover voicing pattern from acoustic detail, that can contribute to the differentiation of words (since voicing is contrastive in English) and also this can make a difference in differentiating the linguistic structure (since the pattern of voicing is influenced by phone position, phrase position, lexical stress, etc.), as phonetic detail can vary based on phrasal context, word frequency, word predictability, individual speaker variation, and indexical information (Pierrehumbert, 2016). This systematic subphonemic variation / fine

phonetic detail should be accounted for in perceptual models because it contains information about the utterance or communicative situation.

Not only may phonetic detail be perceptually useful in human speech, it is also useful in synthetic speech because it contributes to perceptual coherence, which Hawkins and Smith argue is "possibly the main thing that sets [human speech] apart from most synthetic speech" (2001, p. 104). They define perceptual coherence as "what makes a signal robust, natural-sounding, and interpretable as speech" (Hawkins & Smith, 2001, p. 107) and that perceptual coherence is "a central property of natural speech" (Hawkins & Smith, 2001, p. 104). Perceptual coherence has been used in the fields of speech perception and visual perception and the phenomenon is described as follows: "people continuously, without conscious attention, recognize patterns in the stream of sensations that impinge upon them" (Volz & von Cramon, 2006, p. 2077).

However, Hawkins and Smith (2001) also note that a clear definition of what contributes to perceptual coherence is difficult to obtain, as "we do not know exactly what properties make speech perceptually coherent" (2001, p. 108). Some acoustic information thought to contribute to perceptual coherence are patterns of formant frequencies, change in amplitude over time (i.e., amplitude envelope), coarticulatory patterns, and the acoustic events at the boundaries of segments (Hawkins & Smith, 2001, p. 108–9).

In addition to being incorporated into perceptual models, phonetic detail is important to speech technology. For example, the lack of phonetic detail in synthetic speech is argued to increase mental processing costs in speech understanding (Duffy & Pisoni, 1992). When the long-distance effects of the English lateral liquid were incorporated into synthetic speech, listeners better understood the phoneme in noise (Tunley, 1999; West, 1999). Relatedly, these details could be incorporated into models used by forced alignment. Typically, forced alignment has been judged on word error rate (Jurafsky & Martin, 2000, p. 328), but the

boundaries of individual segments are of great importance to phoneticians. The incorporation of phonetic detail into forced alignment will be dealt with further in Chapter 4.

2.4.1 Limitations

While the current study's results show some long-distance coarticulation, there are some limitations in the methodology. First, the use of LPC coefficients is a limitation in that the acoustic correlates of the coefficients are not as clear as those traditionally measured. Taken together, they represent the acoustic signal, but which coefficient corresponds to what traditionally-used acoustic features is less clear. Additionally, state-of-the-art forced alignment systems typically use Mel Frequency Cepstral Coefficients or Perceptual Linear Prediction (Hermansky, 1990) features (which are similar to LPC coefficients, but are supposed to represent what is perceptually important in the acoustic signal), so while this result may not be directly applicable to a state-of-the-art system like Kaldi (Povey et al., 2011), the results do indicate there may be a need to incorporate this fine phonetic detail into acoustic models. Furthermore, as Coleman (2003) says, there is a possibility that imperceptible patterns are being detected, i.e., methods are too sensitive. Additionally, Coleman's study found differences in phones contrasting in word-final voicing. Unfortunately, the distribution of English phones and word-final /t d/ deletion (discussed further in Chapter 4) made it so that in this study, there was only one word with a word-final phone being studied, *look*.

2.5 Conclusion

In the current study, LPC coefficients were measured surrounding stop consonants /p t k b d g/. Results show that these LPC coefficients, at long distance, can be used to predict the voicing of the stop consonant in question at a rate above chance, indicating a relationship between the different realizations of voicing (voiced, bleed, trough, final-third voicing, hump,

and voiceless) and long-distance coarticulation. The existence of this coarticulation has many implications for speech production, speech perception, and speech technology, as this acoustic information can reveal linguistic information to the listener.

Chapter 3

Voice Onset Time in Southern Speech

3.1 Introduction

Sociophonetic variation research traditionally focuses on the quality of vowels, for example, the Southern Vowel Shift (Labov et al., 2006) or the African American Vowel Shift (Thomas, 2001). However, there are some examples in the literature of variation in consonants, for example, "g-dropping", an extensively studied phenomenon in which words with the ending *-ing* are pronounced with the alveolar nasal [n] rather than the velar nasal [ŋ] (Fischer, 1958; Forrest, 2017; Labov, 2006; Yuan & Liberman, 2011), or /p t k/ deletion, a well-studied sociolinguistic phenomenon in which /t/ or /d/ is deleted, especially word-finally in consonant clusters (Bayley, 1994; Guy, 1980; Labov, 1968; Santa Ana, 1991; Stuart-Smith, Sonderegger, Rathcke, & Macdonald, 2015; Tagliamonte & Temple, 2005; Wolfram, 1969)[1]. Another sociolinguistic variation seen in consonants is that of voice onset time. This chapter uses the Digital Archive of Southern Speech (DASS) (Kretzschmar et al., 2013; Kretzschmar et al., 2019) in order to examine variability in voice onset time due to sociolinguistic variables.

[1]This phenomenon will be further addressed in Chapter 4 of this dissertation.

AutoVOT (Keshet et al., 2014) was used to automate measurement of voice onset time, and analysis of the data was performed with a mixed-effects model and boxplots. Results show several linguistic variables, but not sociolinguistic variables, to be significant predictors of VOT. The repetition of the significance of the linguistic variables confirms the replicability of previous studies, and this sociophonetic line of research is important because as we understand sociophonetic differences and lack thereof, we can better understand how we express our social and regional identity. Additionally, this sociolinguistic variation that could be present should be accounted for in speech technology, an idea that is further explored in Chapter 4.

The remainder of this introduction discusses what previous literature asserts about variation in Southern speech (subsection 3.1.1), including prior analyses of DASS; examines what is thought to influence VOT (subsection 3.1.2), including linguistic, speaker-specific, and sociolinguistic variation; and concludes with the research questions raised in light of this prior work and studied in this chapter (subsection 3.1.3).

3.1.1 Variation in Southern Speech

Labov et al. (2006) detail several phonetic properties of Southern speech. In consonants, Southern speech exhibits the presence of /ɹ/ in syllable-final position and g-dropping. However, vowels of Southern speech were more extensively studied, and Labov et al. (2006) report that Southern speech has upgliding, where /æ/ becomes /æy/ before sibilants and nasals, and also front gliding, where /ʊ/ becomes /yʊ/ after coronal onsets. Southern speech also features the fronting of /ʊ/, /u/, /oʊ/, and /æw/. Furthermore, there are mergers and distinctions that make Southern speech unique. There are several mergers: the *pin-pen* merger, where /ɪ/ merges with /ɛ/ before nasals, the *pool-pull* merger, where /ʊ/ and /u/ merge before /l/, the *feel-fill* merger, where /ɪ/ and /i/ merge before /l/, the *fail-fell* merger, where /ɛ/ and /e/ merge before /l/. Additionally, Southern speech distinguishes between /ɔɹ/ and /oɹ/ as in *horse* and *hoarse* respectively, /ʍ/ and /w/ in *which* and *witch* respectively, /eɹV/

and /æɪV/ in *merry* and *marry* respectively, and /ɑ/ and /ɔ/ in *hock* and *hawk* respectively. Southern speech also features /aɪ/ monophthongization thought to be the impetus for the Southern Shift, represented in Figure 3.1.

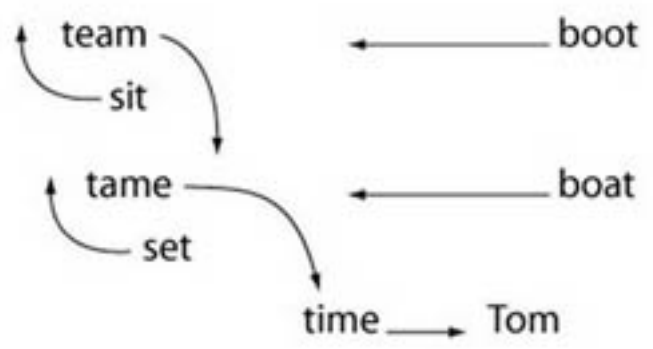

Figure 3.1: The Southern Shift reproduced from Gordon (2005, n. p.)

The present study focuses on the Digital Archive of Southern speech (DASS), and previous analyses of DASS have further shown the complexity of Southern speech: sub-region, phonological environment, sex, and social class affect /aɪ/ weakening (Olsen, Olsen, & Renwick, 2017), differing social factors produce variable implementation of /aɪ/ monophthongization (Renwick & Stanley, 2017), individual variation exists in Southern phonological features (Renwick & Olsen, 2015), and Southern Non-Black speakers and Black speakers' vowel shifts differ (Renwick & Olsen, 2017). An interactive version of this data can be found in the Gazetteer of Southern Vowels (Stanley, Kretzschmar, Renwick, Olsen, & Olsen, 2017). However, these analyses have focused on finding variation in vowels. Only one prior study on consonants in DASS has been done. Jones and Renwick (2020) found evidence of high rates of g-dropping among Southern speakers. These results, along with geographical representations of the mergers, distinctions, and vowel features typical of Southern speech are represented on the GIS Analysis of the Digital Archive of Southern Speech (Jones & Renwick, 2020).

3.1.2 Voice Onset Time

One way that stop consonants may vary is in voice onset time (hereafter, VOT), which is defined as the time from the release of the stop consonant until modal voicing begins (Reetz & Jongman, 2011)[2]. VOT is generally classified into three categories based on length: prevoicing, short-lag (i.e., short voicing lag), and long-lag (i.e., long voicing lag). Prevoicing (or negative VOT) consists of any value of VOT that is negative, typically anywhere from −150 ms to 0 ms (Berry, 2004). The temporal range for short-lag is less clear. Liberman and Blumstein (1988) define short-lag VOT as 0–15 ms, Reetz and Jongman (2011) define it as 0–20 ms, and Berry (2004) defines it as 0–30 ms. Liberman and Blumstein (1988), however define long-lag VOT as 30 ms or longer. To summarize, phonetically voiced stops are prevoiced, phonetically voiceless stops have short-lag VOT, and phonetically voiceless aspirated stops have long-lag VOT.

In English, the phonetic realization of VOT is different for phonologically voiceless and voiced stops. For phonologically voiceless stops /p t k/ (also known as "fortis"), VOT is reported as having long-lag VOT, while phonologically voiced stops /b d ɡ/ (also known as "lenis") are typically reported as having short-lag VOT (Lisker & Abramson, 1964; Smith, 1978). However, Westbury (1979) and Flege and Brown Jr. (1982) report some instances of prevoicing in lenis stops. Jacewicz et al. (2009) suggests that the closure of /b/ in North Carolina English was voiced more often in comparison to Wisconsin speakers. Furthermore, a more recent study of Southern American English found 77.8% of lenis stops to have prevoicing (Hunnicutt & Morris, 2016). As Yao (2009) acknowledges, there is a great deal of variability in the production of VOT, with linguistic, speaker-specific, and social variables having an effect.

[2]Also see Chapter 2.

3.1.2.1 Variation in Voice Onset Time

VOT is well-known to exhibit a great deal of variation, which can be linguistic in nature, where variation is due to place of articulation, stress, etc.; speaker-specific, where variation is a result of physical properties of the speaker, individual differences, etc.; or sociolinguistically, where variation is due to social factors such as age, sex, etc.

3.1.2.1.1 Linguistic Variation in VOT There are many linguistic factors affecting the length of VOT. VOT of voiceless stops is longer before high and tense vowels (Klatt, 1975), and Weismer (1979) also found that VOT is longer before tense vowels. Additionally, Weismer (1979) found that in CVC syllables, the second consonant being voiced increased VOT of the first consonant. More dorsal articulations have longer VOT (Cho & Ladefoged, 1999): it is well attested that the VOT of [k^h] is longer than [p^h] (Klatt, 1975; Peterson & Lehiste, 1960), but the reported duration of VOT of [t^h] with respect to [p^h] and [k^h] is less consistent. Most studies find the VOT of [t^h] is roughly equivalent to [k^h] (Docherty & Ladd, 1992; Yao, 2009), but other studies report values of VOT of [t^h] between that of [p^h] and [k^h] (Lisker & Abramson, 1964; Peterson & Lehiste, 1960). Additionally, monosyllabic words have longer VOT than disyllabic words (Flege, Frieda, Walley, & Randazza, 1998). Another linguistic variable is word frequency. According to Yao (2009), words with a higher frequency have a shorter VOT. However, Flege et al. (1998) suggest that word *familiarity*, rather than frequency, plays a role in VOT duration. Additionally, stress plays a role in VOT, with voiceless stops in stressed syllables having a longer VOT (Lisker & Abramson, 1967).

3.1.2.1.2 Speaker-specific Variation in VOT There is a great deal of variation in VOT; for example, we know that VOT varies cross-linguistically (Cho & Ladefoged, 1999) and due to influence from knowledge of other languages (Nagy & Kochetov, 2013). Additionally, there are physiological factors that affect VOT. Speakers with a high lung volume have longer VOTs, while people with lower lung volumes have shorter VOTs (Hoit, Solomon, & Hixon,

1993). Even when controlling for speaking rate, VOT features a great deal of speaker-specific variation: "Variability across talkers, particularly among the voiceless categories, can span tens of milliseconds, making this source one of the larger factors in VOT variation" (Chodroff & Wilson, 2017, p. 31). However, as Chodroff and Wilson (2017) point out, the duration of VOT varies in a predictable way: "While the realization of speech sounds is highly variable, it is also highly patterned or structured" (2017, p. 30), and VOT length varies systematically within a talker's speech.

3.1.2.1.3 Sociolinguistic Variation in VOT In addition to the linguistic and speaker-specific variation in VOT, there is also variation that results from sociolinguistic variables. In the study "Variation in Voice Onset Time Along the Scottish-English Border", Docherty, Watt, Llamas, Hall, and Nycz (2011) found speakers' country and coast (west vs. east) to be significant social factors that predict length of VOT. Shetland English speakers' VOT varied based on the dialect of their parents (English, Shetland English, or Scottish English) (Scobbie, 2005). Additionally, Stuart-Smith et al. (2015) found expected effects of place of articulation and speech rate in Glaswegian English.

As previously mentioned, there is some evidence for differences in VOT of American English dialects, as Hunnicutt and Morris (2016) suggest that Southern American English lenis stops are more likely to have negative VOT than short-lag VOT, which is in line with Jacewicz et al. (2009), who found that Southern speakers were more likely to have voicing in the closure of [b]. Additionally, Walker (2020)[3] found that actors asked to imitate a Southern accent produced more negative VOT stops, meaning that voiced stops are a "marker" (Labov, 1972) of Southern American English.

Walker's (2020) study also can be taken to support the "laryngeal realism" approach (Honeybone, 2005). Honeybone (2005) posits that the contrast in English stops typically represented by the feature [voice] is better described as the feature [spread glottis] (Halle

[3]It should be acknowledged that Walker (2019) served as the inspiration for the present study.

& Stevens, 1971) in an approach termed "laryngeal realism". Davidson (2016) found that the voicing of phonologically voiced obstruents depends a great deal on articulatory and aerodynamic factors, and Davidson (2018) finds this to be somewhat true of voiceless obstruents, though effects are "tempered by the requirements that are imposed by a laryngeal abduction gesture"(2018, p. 351). In Southern English, it is possible that there is some laryngeal mechanism that preserves voicing:

> Given that [North Carolina (NC) speakers'] closures were mostly fully voiced and the proportion of voicing during the closure was generally insensitive to the variation in closure length as a function of word emphasis, it appears that NC speakers maintained transglottal pressure during the stop closure by **some active articulatory maneuvers** [emphasis mine], which may or may not be the same as for the [Wisconsin] speakers (Jacewicz et al., 2009).

In addition to geographical factors, other social factors were found to affect VOT. The effect of sex on the duration of VOT is somewhat contested. Swartz (1992) found that men have shorter VOT than women for /t/ and /d/, while Whiteside and Irving (1997, 1998) found that sex interacted with voicing of the stop: women have longer VOT for voiceless stops and shorter for voiced stops. However, a more recent study has found no difference in sex (Morris, McCrea, & Herring, 2008). Additionally, Docherty et al. (2011) found speakers' age to predict VOT: younger speakers had a longer VOT for voiced stops and shorter VOTs for voiceless stops.

Very few studies have dealt with ethnicity's effect on voice onset time. Ryalls, Zipprer, and Baldauff (1997) found an effect of gender and race, with African Americans having larger negative VOTs than Caucasian Americans. They did find an interaction between ethnicity and voicing, as African Americans only exhibited differences in the phonologically voiced stop consonants. However, when the first author tried to replicate the findings of this study, Ryalls, Simon, and Thomason (2004) were unable to find an effect for race.

3.1.3 Research Questions

Due to a lack of reproducible effect for ethnicity (Ryalls et al., 2004; Ryalls et al., 1997), contradicting effects of sex in the prior literature (Morris et al., 2008; Whiteside & Irving, 1997, 1998), and limited research on age's effect on VOT, a new look at VOT in Southern speech is needed. The DASS corpus is an excellent corpus from which to collect Southern speaker data due to the extensive processing that has been done and due to the highly naturalistic speech it contains. Additionally, DASS data offers a historical look at VOT in Southern speech, which could then be used for comparison to modern speech. My research question is: How do sociolinguistic variables, ethnicity, sex, age, in combination with linguistic variables, such as stress of the following vowel and place of articulation of the stop, affect the realization of VOT in the Digital Archive of Southern Speech (Kretzschmar et al., 2013; Kretzschmar et al., 2019)? My hypotheses are as follows: 1. effects for ethnicity will be found, where Black speakers will have increased VOT while Non-Black speakers will have short VOT, following Ryalls et al. (1997), 2. effects for sex will be found in both voiced and voiceless data, following Whiteside and Irving (1997, 1998), where women will have shorter positive VOT, and 3. effects for age will be found, with younger speakers will have longer VOTs for voiced stops and shorter VOTs for voiceless stops, following Docherty et al. (2011).

3.2 Methodology

This methodology section first describes the corpus studied, the Digital Archive of Southern Speech in subsection 3.2.1, followed by the software used to automatically extract values of VOT in subsection 3.2.2, and concludes with the variables measured in subsection 3.2.3.

3.2.1 Digital Archive of Southern Speech

To study variation of VOT in Southern speech, the Digital Archive of Southern Speech (DASS) (Kretzschmar et al., 2013; Kretzschmar et al., 2019) was used. DASS is a representative sample of the Linguistic Atlas of the Gulf States (consisting of eight Gulf States) (Pederson, McDaniel, & Adams, 1986), itself a subset of the larger Linguistic Atlas Project. DASS is a fully transcribed historical audio corpus that was recorded from 1970 to 1983 and contains 367 hours of audio. There are 64 speakers: 34 male and 30 females, born in years 1886 to 1965, with a mean age of 61 years of age. DASS was used due to its novelty and due to the collection of sociolinguistic variables from the speakers. In the processing of the corpus, sound files were force aligned with the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). The forced alignments were validated by listening to a sample of 2115 words in the corpus, and the word boundaries were found to be 86.7% accurate (92% accurate counting partially accurate tokens)[4].

3.2.2 AutoVOT

AutoVOT (Keshet et al., 2014) is software consisting of an algorithm that is used for automatic measurement of positive VOT of word-initial prevocalic stops, having been trained on human-labeled VOT measurement. Unfortunately, AutoVOT is unable to mark negative VOT; however, previous studies have utilized AutoVOT in order to measure positive VOT of phonologically voiced stops (Chodroff & Wilson, 2017; Xuanda, Ziyu, & Jian, 2018). The software requires wav files and time intervals in which to look for VOT, and the software includes pre-trained classifiers. A Praat (Boersma & Weenink, 2018) script was used to find instances where the left boundary of /p, t, k, b, d, g/ coincided with the start time of a word so long as the phone was also followed by a vowel. In this same Praat script, new TextGrids were

[4]Accuracy rates determined by Keiko Bridwell, a research assistant in the Linguistic Atlas Project. Bridwell and Renwick have listened to approximately 10,000 additional tokens and have found comparable rates in that subset.

created that marked windows of time, extending forced alignment boundaries, for AutoVOT to search. These boundaries were 31 ms in both directions for voiceless stops and 11 ms in both directions for voiced stops, following Chodroff and Wilson (2017). The pre-trained Amanda classifier provided by AutoVOT was used. AutoVOT was not retrained, following Chodroff and Wilson (2017), who found that training on the corpus at hand did not perform as well as using the pre-trained model: "We trained on two-thirds of our manually-measured stops (1488 voiceless, 990 voiced) and tested on the remaining third (755 voiceless, 489 voiced). The root-mean-square deviation of the resulting model (13.0 ms) was not superior to that of the pretrained models" (2017, p. 39).

3.2.3 Variables

Linguistic variables (detailed in subsubsection 3.2.3.1) were measured and sociolinguistic variables from DASS were recorded. Linguistic variables include following vowel height, following vowel tenseness, syllable stress, word syllable count, frequency, and speaking rate (measured by duration of the following vowel). Sociolinguistic variables (detailed in subsubsection 3.2.3.2) include sex, age level, and ethnicity.

3.2.3.1 Linguistic Variables

This section details the specifics of how following vowel height, following vowel tenseness, syllable stress, word syllable count, frequency, and speaking rate were dealt with in this study. Vowel height consists of three levels: high ([ɪ, i, ʊ, u]), mid ([ʌ, ɔ, ɛ, ɝ, eɪ, oʊ, ɔɪ]), and low ([ɑ, æ, aɪ, aʊ]). Vowel tenseness consists of two levels: tense ([ɑ, aʊ, aɪ, eɪ, i, oʊ, ɔɪ, u]) and not tense ([ʌ, ɔ, ɛ, ɝ, ɪ, ʊ]). Klatt (1975) found an effect of following vowel height and tenseness for *voiceless* stops, but these linguistic variables were also used in the voiced analysis. Place of articulation of the stop consisted of three levels: labial (/p, b/), coronal (/t, d/), and dorsal (/k, g/), as more dorsal places of articulation feature longer VOTs in

voiceless stops (Cho & Ladefoged, 1999). As Klatt (1975) found stress of voiceless stops to affect VOT, stress of the syllable containing the stop as well as number of syllables in the word was classified from pronunciation output by MFA, which uses a dictionary based on the LibriSpeech Corpus. To control for effects of frequency, a list of stop words used in previous DASS analyses (listed in Appendix B) were removed. In order to control for speaking rate, the duration of the following vowel was used. There are other ways of measuring speech rate such as syllable duration, but as Theodore, Miller, and DeSteno (2009) point out, syllable duration is correlated with duration of VOT because the stop is contained within the syllable duration. Additionally, according to Theodore et al. (2009), the speaking rate affects VOT duration, but it affects different people at different rates. Linguistic variables are summarized in Table 3.1.

Table 3.1: Linguistic variables

Variable	**Values**
Vowel height	High
	Mid
	Low
Tenseness	Tense
	Lax
Place of articulation	Labial
	Coronal
	Dorsal
Stress	Primary
	Secondary
	Unstressed
Vowel duration	Numeric Variable

3.2.3.2 Sociolinguistic Variables

Sociolinguistic variables relevant to the current study that were recorded in DASS are summarized in Table 3.2 and Table 3.3, and include sex (male or female), age level (13–45, 46–65, 66–76, 77–99), and ethnicity (Black or Non-Black).

Table 3.2: Sociolinguistic variables

Variable	**Values**	**Number of participants**
Sex	Male	33
	Female	31
Age level	13-45 years old	19
	46-65 years old	12
	66-76 years old	16
	77-99 years old	17
Ethnicity	Non-Black	49
	Black	16

Table 3.3: Number of participants by group

Sex	**Ethnicity**	**Age level**	**Number of participants**
Male	Non-Black	13–45	6
		46–65	2
		66–76	9
		77–99	7
	Black	13–45	4
		46–65	1
		66–76	0
		77–99	5
Female	Non-Black	13–45	7
		46–65	8
		66–76	5
		77–99	4
	Black	13–45	2
		46–65	1
		66–76	2
		77–99	1

3.3 Results

Data was first organized using R (R Core Team, 2018) and RStudio (RStudio, 2017) along with R packages 'tidyverse' (Wickham, 2017) and 'plyr' (Wickham & Francois, 2015). Then,

in order to analyze the data, boxplots were created with 'ggplot2' (Wickham, 2016) and mixed models were created using R packages 'lme4' (Bates, Maechler, Bolker, & Walker, 2015) and 'lmerTest' (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2016). Output of models was produced using the R package 'xtable' (Dahl, Scott, Roosen, Magnusson, & Swinton, 2019).

A small number of tokens were omitted due to their windows that were automatically marked overlapping, for example, the 62 ms window of the [t]s "to talk" would overlap. In the event that a stop was transcribed as a word itself (i.e., a stutter), this was removed from analysis. Additionally, stops that AutoVOT reported as "zero confidence" were removed. A total of 109,019 voiceless stops (28,139 instances of [p], 29,889 instances of [t], and 50,986 instances of [k]) and a total of 92,630 voiced stops (37,476 instances of [b], 38,813 instances of [d], and 16,341 instances of [g]) are included in the following analyses. Table 3.4 shows the number of stops broken down by ethnicity.

Table 3.4: Stop count by ethnicity

Ethnicity	**Stop**	**Count**
Non-Black	[b]	27,988
	[d]	28,793
	[g]	11,922
	[p]	20,722
	[t]	22,119
	[k]	37,198
Black	[b]	9488
	[d]	10,020
	[g]	4419
	[p]	7418
	[t]	7776
	[k]	13,786

3.3.1 Voiceless Stop Results

Figure 3.2 shows the raw data in histograms for voiceless stops. The distributions are right-skewed, and show a relatively low number of outliers. For this reason, the outliers are not excluded.

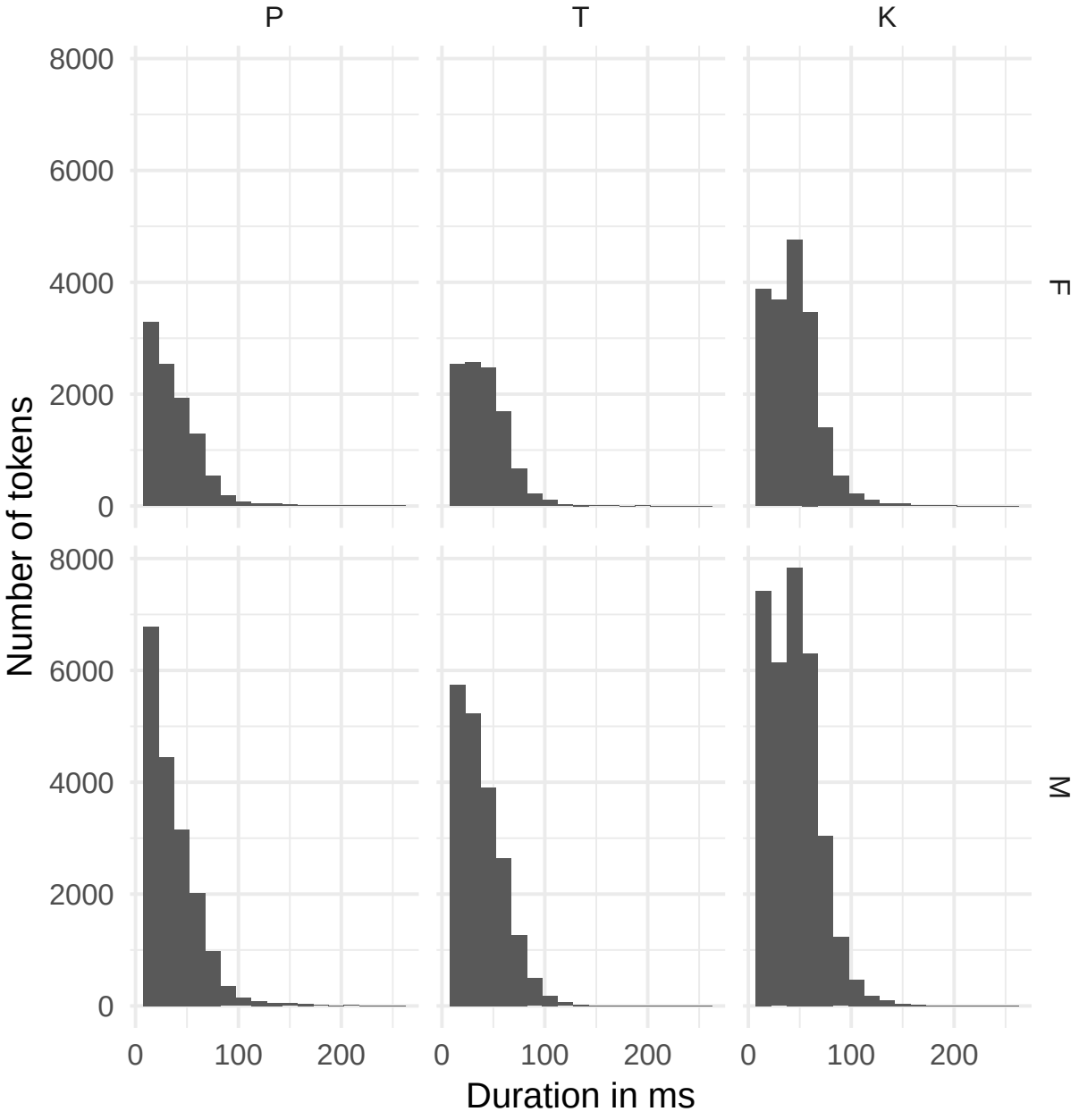

Figure 3.2: Voiceless stop VOT duration data

Mean and standard deviation values are given in Table 3.5. The mean of the stop increases as the place of articulation becomes more dorsal, and each distribution has a similar standard deviation. Summary statistics for each group are shown in Table 3.6.

Table 3.5: Voiceless stop VOT data summary

Sex	**Stop**	**Mean**	**SD**	***n***
Female	[p]	38.1 ms	24.6 ms	9987
	[t]	40.7 ms	21.7 ms	10,308
	[k]	44.2 ms	23.3 ms	18,181
Male	[p]	36.7 ms	24.5 ms	18,155
	[t]	38.8 ms	22.5 ms	19,578
	[k]	44.8 ms	23.9 ms	32,806

To investigate all interactions between the three sociolinguistic variables (age, ethnicity, and sex), boxplots were made that show ethnicity by sex (shown in Figure 3.3), sex and age level (shown in Figure 3.4, and age level and ethnicity (shown in Figure 3.5).

In both Figure 3.3 and Figure 3.4 there does not appear to be an interaction between either ethnicity and sex or age level and sex, though Figure 3.5 shows that for age level 1 (13–45) the duration is longer for Non-Black speakers than for Black speakers, while the reverse is true at age levels 2 (46–65), 3 (66–76), and 4 (77–99). Due to this, an interaction will be included in the mixed model.

To summarize the variables listed above, the formula for the mixed model is given in Equation 3.1.

$$\begin{aligned}\text{Voice onset time duration} \sim\ & \text{vowel duration} + \text{vowel height} + \text{vowel tenseness} + \\ & \text{syllable stress} + \text{stop place of articulation} + \text{sex} + \\ & \text{age level} * \text{ethnicity} + \text{number of syllables} + \\ & (1|\text{word}) + (\text{vowel duration} \,||\, \text{speaker}) \end{aligned} \tag{3.1}$$

Table 3.6: Voiceless stop VOT data by ethnicity summary

Sex	Ethnicity	Stop	Mean	SD	*n*
Female	Non-Black	[p]	37.2 ms	24.4 ms	8237
		[t]	40.1 ms	21.6 ms	8610
		[k]	43.5 ms	23.1 ms	14,921
	Black	[p]	42.3 ms	24.8 ms	1750
		[t]	43.6 ms	21.9 ms	1698
		[k]	47.3 ms	23.8 ms	3260
Male	Non-Black	[p]	35.9 ms	24.5 ms	12,486
		[t]	38.1 ms	22.7 ms	13,504
		[k]	44.3 ms	23.6 ms	22,280
	Black	[p]	38.3 ms	24.5 ms	5669
		[t]	40.4 ms	22.2 ms	6074
		[k]	46.0 ms	24.4 ms	10,526

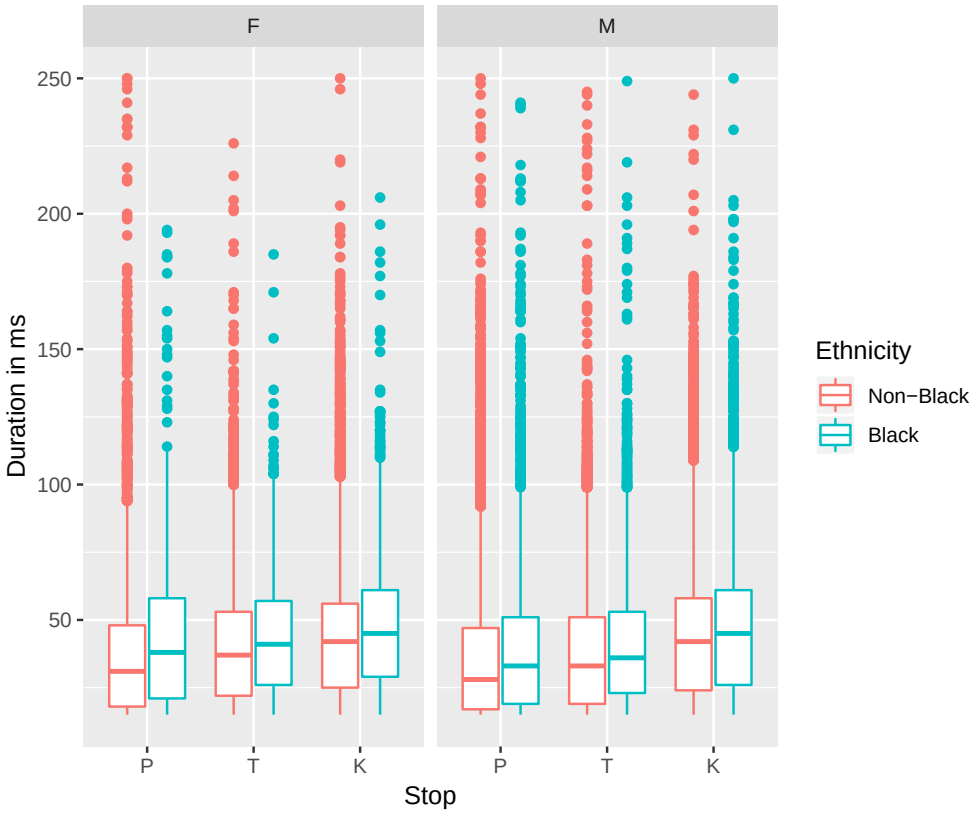

Figure 3.3: Voiceless stop VOT data by ethnicity and sex

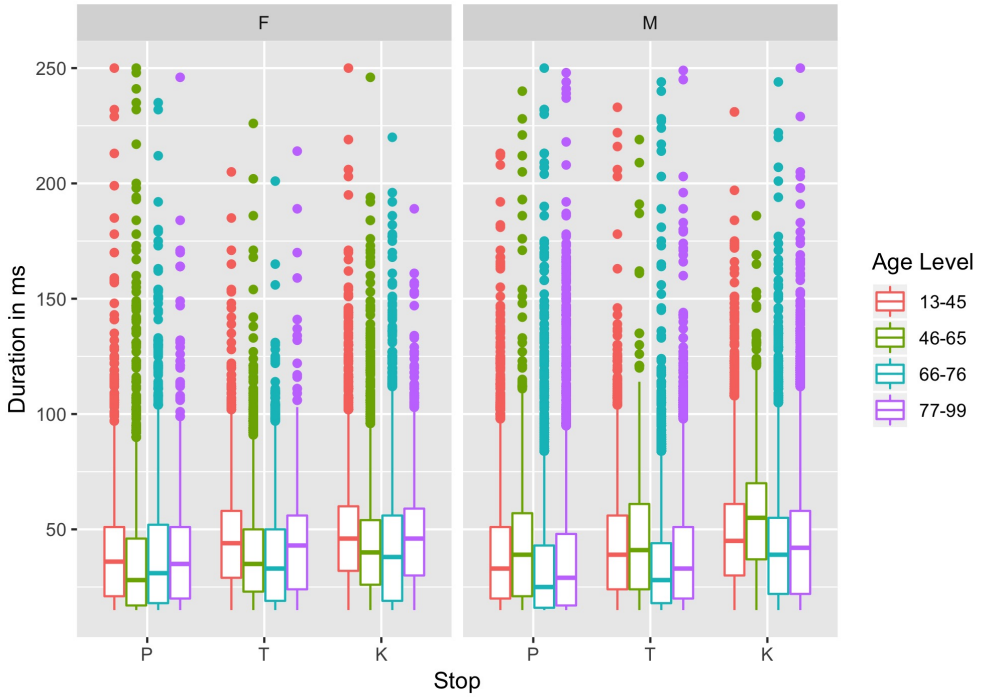

Figure 3.4: Voiceless stop VOT data by age level and sex

In addition to the variables and interactions as outlined in subsubsection 3.2.3.1 and subsubsection 3.2.3.2, 'word' and 'speaker' were added as random factors. An uncorrelated random slope was added for speaker by vowel duration, in order to account for the fact that speaking rate (here, represented by following vowel duration) has a different effect on VOT for different speakers (Theodore et al., 2009). The results of this mixed model are given in Table 3.7.

As we can see from the mixed model, vowel duration has a significant effect ($p = 0.00$) on VOT: as vowel duration increases by 1 s, VOT increases by 37 ms. The VOT of a voiceless stop before a mid vowel height was significantly ($p = 0.00$) shorter, 3.5 ms, than VOT preceding a high vowel, which is what we would expect according to Klatt (1975), and the VOT of a voiceless stop before a low vowel height was also significantly ($p = 0.00$) shorter than VOT of a stop preceding a high vowel, 2.3 ms. The vowel following the stop

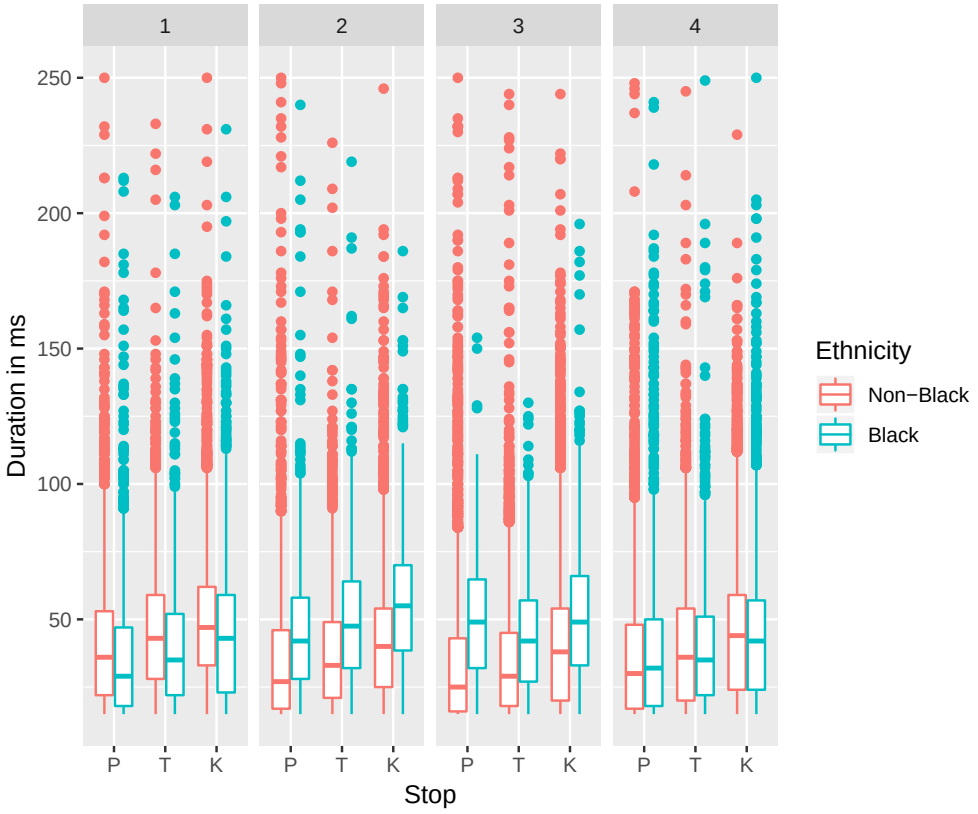

Figure 3.5: Voiceless stop VOT data by ethnicity and age level

consonant being tense was marginally significant ($p = 0.05$) with a small effect size; the VOT of stops followed by tense vowels is 0.76 ms longer than that of lax vowels. Additionally, all stress levels had a statistically significant effect ($p = 0.00$), both primary and secondary stress increased VOT, by 7.4 and 6.6 ms respectively, which would be expected because stress strengthens segments (and aspiration increases the strength of a segment) (Fougeron & Keating, 1997). Both coronal and dorsal places of articulation were significantly ($p = 0.00$) different than labial, coronal stops being longer than labial stops by 4.2 ms, and dorsal stops being longer than coronal stops by 10.1 ms, in line with the research of Peterson and Lehiste (1960) and Lisker and Abramson (1964). The third age level, 66–76, was marginally

significantly ($p = 0.02$) shorter than the reference age level 13–45 by 4.5 ms, but there was no interaction between age and ethnicity. Sex, ethnicity, and number of syllables had no effect.

Table 3.7: Mixed Model for Voiceless Stops

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	30.56	1.75	92.07	17.43	0.00
vowel duration (ms)	36.77	3.94	67.02	9.32	0.00 (***)
vowel height = low	-3.48	0.53	1541.38	-6.62	0.00 (***)
vowel height = mid	-2.32	0.49	1449.71	-4.76	0.00 (***)
tense	0.76	0.39	1667.62	1.95	0.05 (*)
primary stress	7.37	0.65	2501.77	11.32	0.00 (***)
secondary stress	6.55	1.28	2919.01	5.13	0.00 (***)
coronal	4.16	0.49	1325.42	8.43	0.00 (***)
dorsal	10.11	0.43	1391.92	23.54	0.00 (***)
male	2.01	1.38	53.96	1.46	0.15
age level = 46–65	-2.83	2.16	54.96	-1.31	0.19
age level = 66–76	-4.53	1.95	53.63	-2.32	0.02 (*)
age level = 77–99	-1.41	2.07	53.85	-0.68	0.50
Black	0.68	2.51	54.85	0.27	0.79
disyllabic	0.39	0.42	1213.55	0.93	0.35
polysyllabic	-0.06	0.59	2068.49	-0.11	0.91
age level = 46–65:Black	6.35	4.59	52.74	1.38	0.17
age level = 66–76:Black	8.50	4.69	54.10	1.81	0.08
age level = 77–99:Black	-1.25	3.55	53.59	-0.35	0.73

3.3.2 Voiced Stop Results

Figure 3.6 shows the raw data for voiced stops. The data is right skewed, and like the voiceless stop data, there are few outliers, which are included in the data.

The mean and standard deviation values are given in Table 3.8, and summary statistics for each group are shown in Table 3.9.

Boxplots similar to those shown in the voiceless data, which show three sociolinguistic variables (age, ethnicity, and sex), are used to visualize any interactions that may be present. They show ethnicity by sex (shown in Figure 3.7), sex and age level (shown in Figure 3.8, and age level and ethnicity (shown in Figure 3.9).

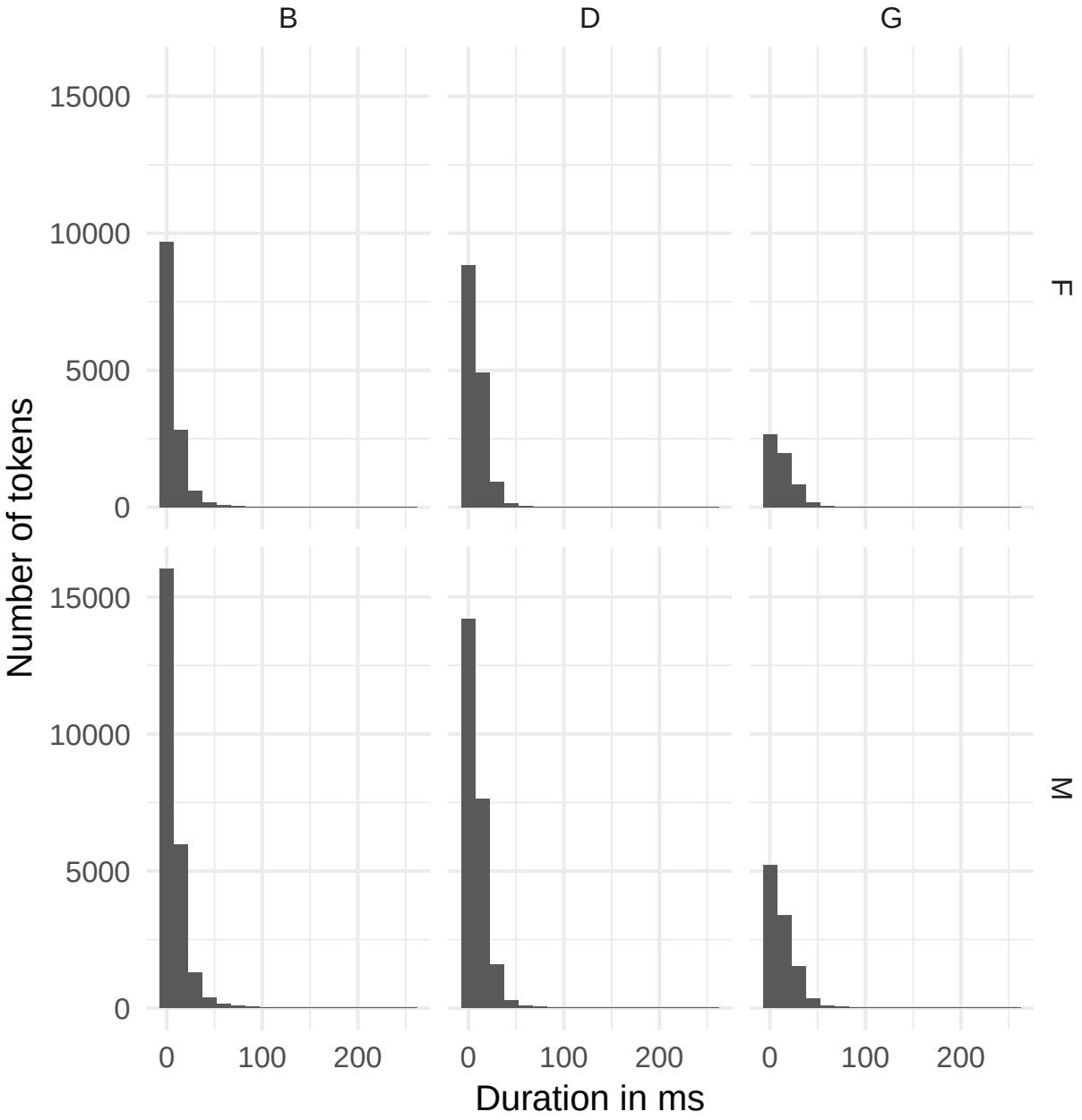

Figure 3.6: Voiced stop VOT duration data

Figure 3.7 and Figure 3.8 do not suggest any interactions. However, in Figure 3.9, there may be an an interaction between age level and ethnicity, as the difference in distribution between Black speakers and Non-Black speakers for age level 1 (13–45) trends in the opposite

Table 3.8: Voiced stop VOT data summary

Sex	**Stop**	**Mean**	**SD**	***n***
Female	[b]	8.88 ms	12.6 ms	13,434
	[d]	9.74 ms	10.0 ms	14,927
	[g]	13.7 ms	14.0 ms	5745
Male	[b]	9.98 ms	15.3 ms	24,042
	[d]	9.91 ms	10.9 ms	23,886
	[g]	13.4 ms	14.0 ms	10,596

Table 3.9: Voiced stop VOT data by ethnicity summary

Sex	**Ethnicity**	**Stop**	**Mean**	**SD**	***n***
Female	Non-Black	[b]	8.53 ms	12.3 ms	11,245
		[d]	9.54 ms	10.2 ms	12,216
		[g]	13.3 ms	13.9 ms	4682
	Black	[b]	10.7 ms	13.8 ms	2189
		[d]	10.6 ms	9.48 ms	2711
		[g]	15.7 ms	14.5 ms	1063
Male	Non-Black	[b]	9.55 ms	14.5 ms	16,743
		[d]	9.70 ms	14.4 ms	16,577
		[g]	13.5 ms	13.8 ms	7240
	Black	[b]	11.0 ms	17.0 ms	7299
		[d]	10.4 ms	11.2 ms	7309
		[g]	13.2 ms	14.5 ms	3356

direction than the other age groups. For this reason, an interaction between age level and ethnicity is included in the mixed model.

In the case of voiced VOT, the results of the mixed model shown in Table 3.10 give somewhat similar results. Vowel duration was found to have a marginally significant ($p = 0.02$) effect; as vowel duration increases by 1 s, VOT increases by 2.4 ms, a much smaller effect than found in voiceless stops. Unlike voiceless stops, vowel height is not a significant factor. Tense is also significant ($p = 0.00$), with stops preceding tense vowels showing a decrease in VOT of 0.8 ms. Only primary stress had a significant effect ($p = 0.00$), increasing positive

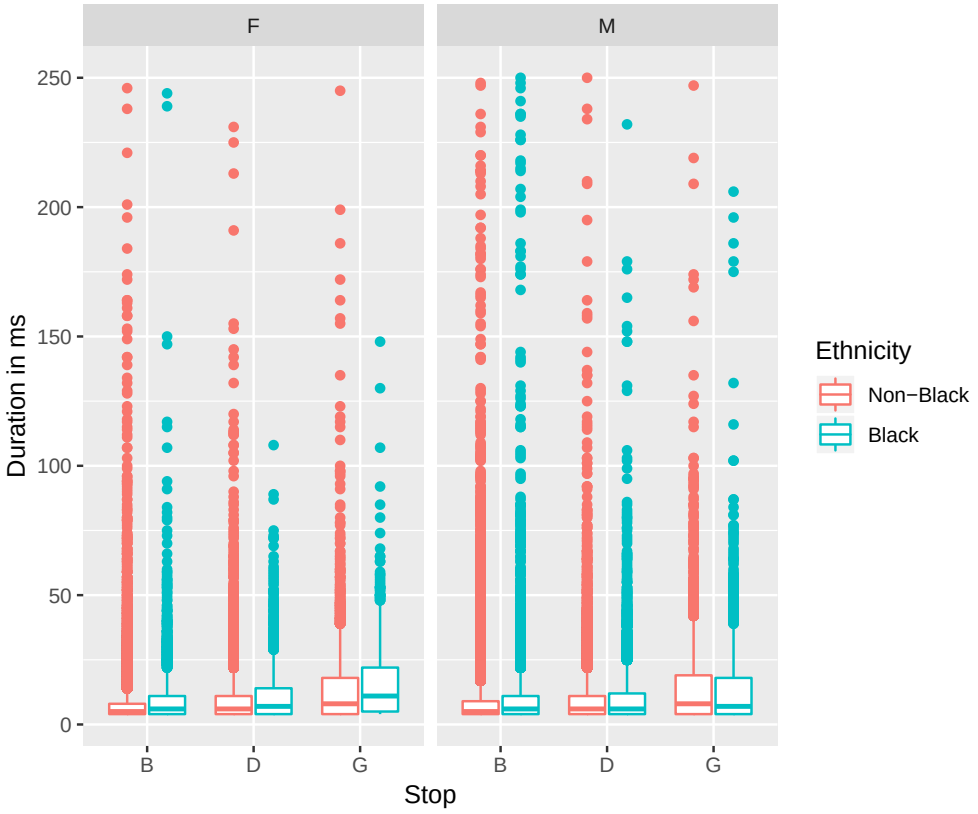

Figure 3.7: Voiced stop VOT data by ethnicity and sex

VOT by 1.6 ms. Coronal stops and dorsal stops were significantly ($p = 0.02$ and $p = 0$, respectively) longer than labial stops, by 0.7 and 5.9 ms respectively, and the effect is much greater for dorsal stops. Sex, age level, ethnicity, number of syllables, and the interaction between sex and age level were not significant.

3.4 Discussion

My hypothesis that ethnicity would have an effect on the length of VOT is not supported by the statistical models, and this follows Ryalls et al.'s (2004) conclusion. However, there does at least seem to be a trend that Black speakers have longer VOT than Non-Black speakers,

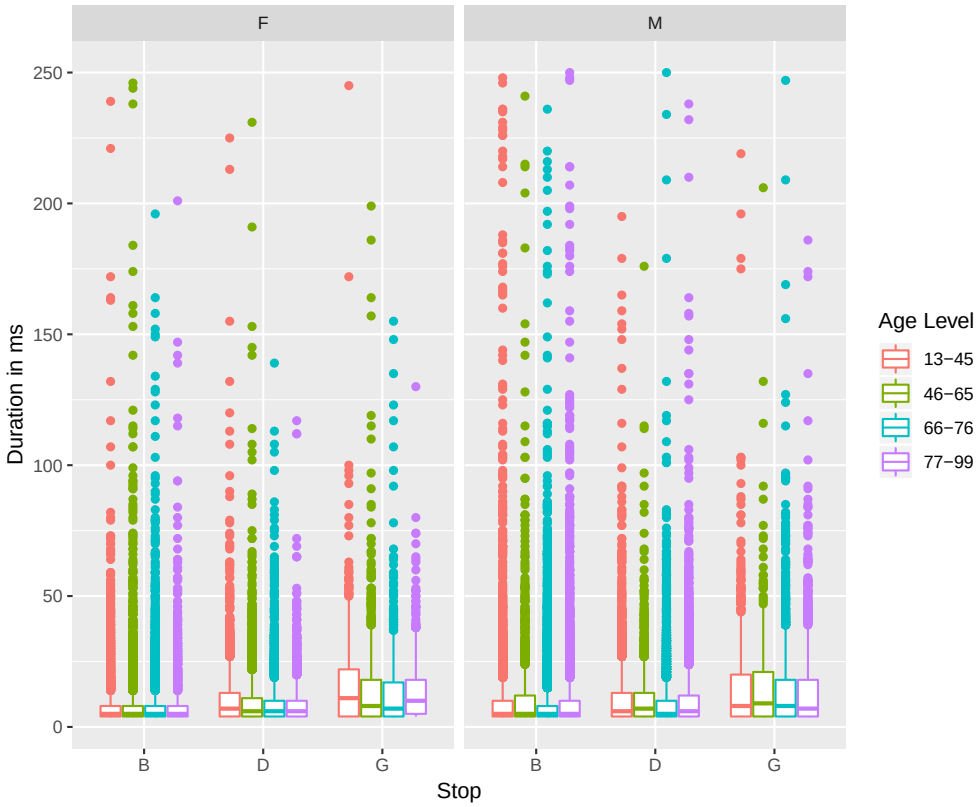

Figure 3.8: Voiced stop VOT data by age level and sex

indicated by boxplots in Figure 3.7 and Figure 3.3, in line with Ryalls et al. (1997). The second part of my other hypothesis, that women would have longer positive VOT than men, is not supported by the data, and this follows Morris et al.'s (2008) findings. Finally, there may be an effect of age in the voiceless data, but the effect is only present for one age group, 66–76. These findings are also limited by the methodology; negative VOT was not measured, and it is possible that differences exist in the closure portion of the stops. This and other limitations are further discussed in subsection 3.4.1.

For the voiceless data, it is unsurprising to see effects of the linguistic variables, vowel duration, vowel height, stress, and place of articulation. Only one age level had a significant effect, contrary to Docherty et al.'s (2011) Scottish English study, though of course findings

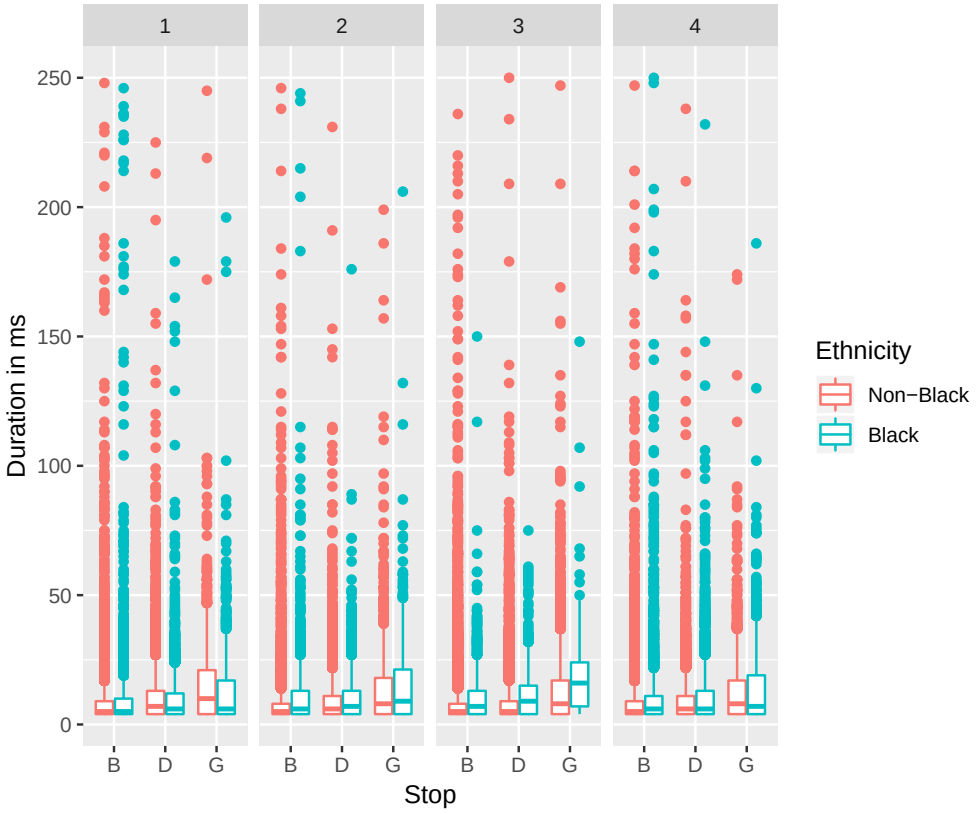

Figure 3.9: Voiced stop data by ethnicity and age

from Scottish English may not be applicable to Southern American English speech. This finding could also be related to the fact that older speakers comprise DASS participants. Number of syllables was also not found to have an effect, contradicting Flege et al. (1998).

For voiced data, the results are not very different. Most of the linguistic factors that were significant in voiceless VOT are significant in voiced VOT. Vowel duration was significant, but with a much smaller effect size than in voiceless VOT. Also, vowel height was not significant in the voiced data, while it was in the voiceless data. Tense was significant in the voiced data and only marginally significant in the voiceless data. For voiced data, stops in syllables that have primary stress had a longer VOT than unstressed syllables, but there was no effect for secondary stress. Coronal stops are slightly longer than labial stops, and dorsal stops

Table 3.10: Mixed model for voiced stops

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	8.36	0.73	123.82	11.43	0.00 (***)
vowel duration (ms)	2.38	0.99	81.02	2.40	0.02 (*)
vowel height = low	0.43	0.36	788.86	1.21	0.23
vowel height = mid	0.42	0.30	848.56	1.39	0.16
tense	-0.82	0.27	804.90	-3.01	0.00 (***)
primary stress	1.61	0.34	2909.00	4.68	0.00 (***)
secondary stress	0.29	0.80	3736.53	0.36	0.72
coronal	0.71	0.29	604.72	2.40	0.02 (*)
dorsal	5.85	0.40	540.55	14.74	0.00 (***)
male	0.80	0.54	54.98	1.49	0.14
age level = 46–65	0.01	0.84	55.99	0.01	0.99
age level = 66–76	-0.86	0.76	54.21	-1.14	0.26
age level = 77–99	-0.78	0.81	54.87	-0.97	0.33
Black	-0.03	0.97	55.56	-0.03	0.98
disyllabic	-0.31	0.25	1108.06	-1.28	0.20
polysyllabic	-0.28	0.38	1506.26	-0.74	0.46
age level = 46–65:Black	1.33	1.78	53.42	0.74	0.46
age level = 66–76: Black	3.42	1.82	55.41	1.87	0.07
age level = 77–99: Black	1.72	1.38	54.48	1.25	0.22

are much longer (almost 6 ms) than labial stops. The three sociolinguistic variables, age, ethnicity, and sex (and the interaction between age level and ethnicity) are not significant.

It is unsurprising that the effect sizes for the voiced stop model are small, as the measurement of positive VOT resulted in small numbers (see Table 3.8) as opposed to voiceless VOT (see Table 3.5). There are fewer predictors in the voiced model than the voiceless model; the voiced model did not have significant predictors for secondary stress and vowel height. The fact that vowel height is not significant for voiced stops is unsurprising, as it is in line with the results of Hunnicutt and Morris (2016), perhaps due to the fact that these stops may be prevoiced, where the voicing is further away in time from the subsequent vowel. As for why there would be an effect of primary stress and not secondary, this could be due to a weaker

effect of stress in general, so that the difference is only distinguishable between unstressed (the reference level) and primary stress.

3.4.1 Limitations

There are several limitations in this study due to measurement and nature of the variable being measured. After the data processing of this study had been completed, a new software, Dr. VOT (Shrem, Goldrick, & Keshet, 2019) was released. In future work, it may be beneficial to use this software, as Dr. VOT can measure negative as well as positive VOT. This may be especially useful for Southern speech, as Hunnicutt and Morris (2016) finds Southern speech to have negative VOT in phonologically voiced stops. While this study only measured positive VOT, an examination of negative VOT would allow further investigation into articulatory nature of (i.e., the state of the larynx) Southern English stops.

Additionally, due to the nature of VOT, it is difficult to capture all sources of variation. One additional limitation of the results is that small differences may not be perceptually relevant, or may not be above the threshold of just-noticeable difference (Klatt, 1976; Klatt & Cooper, 1975) and are more appropriate as evidence in a physiological explanation of acoustic differences.

3.5 Conclusion

This chapter investigates VOT in the DASS (Kretzschmar et al., 2013; Kretzschmar et al., 2019) corpus. For voiceless stops, linguistic variables that we would expect to have an effect (vowel duration, vowel height, stress, and place of articulation) do. For voiced stops, linguistic variables that have an effect are vowel duration, tense, stress (though only primary, in comparison with unstressed syllables), and place of articulation have an effect on positive VOT. This study finds a relationship between the linguistic variables and VOT, but fails

to find a relationship between the sociolinguistic variables age, ethnicity, and sex, and VOT. Furthermore, the linguistic variables studied affect voiced and voiceless stops differently.

3.6 Acknowledgements

This research was supported by NSF BCS #1625680 and the American Dialect Society.

Chapter 4

Automatic Detection of /t d/ Deletion using Forced Alignment

4.1 Introduction

In phonetic research, the time required to manually annotate large-scale data is often prohibitive, and computer automation is needed. Additionally, manual annotation is prone to error from human perceptual systems that may "reconstruct" pronunciation variants to their canonical forms (Mitterer, 2011). One such technique for automated annotation is forced alignment, an offshoot of automatic speech recognition, which provides word and phone boundaries of use to linguists. However, forced alignment systems rely on a dictionary that typically gives canonical pronunciations of words, which is a problem for any kind of variation at the level of the phone (e.g., "old boy" /oʊld bɔɪ/ → [oʊl bɔɪ])[1]. This chapter investigates the

[1]The focus of this dissertation that is hopefully obvious by now is *fine phonetic detail*. One could say that change at the level of the phone is no longer *fine phonetic detail*. However, *any* kind of categorization (as opposed to measurement of a continuous numeric variable) removes us further from fine phonetic detail, a shortcoming pointed out in subsection 4.4.4. It is worth pointing out that, in some way, the entire field of acoustic measurement can only get *categorical* snapshots in time, as audio sampling takes measurements

efficacy of modifying the dictionary of the Montreal Forced Aligner (McAuliffe et al., 2017) to account for a well-attested sociophonetic variation phenomenon, /t d/ deletion, by aligning the Buckeye Corpus (Kiesling et al., 2006), which was chosen as it allows comparison of forced alignment results to human transcriber results. Overall, 23,522 tokens were examined. Forced alignment results from this modification were in agreement with human transcribers approximately 71% of the time, close to the 76% agreement of human transcribers (Raymond et al., 2002). These results are promising for future large-scale sociophonetic research in which dictionary modifications can be made to better force align data containing sociophonetic variation. Additionally, there was an examination of fine phonetic detail at the subphonemic level, as 10% of the corpus was examined manually to explore subphonemic properties of /t d/ and how these properties related to MFA's judgment of the presence or absence of [t, d]. Results show that for /d/, the presence of a stop burst was statistically significant in MFA's determination. For /t/, closure, burst, delayed release, glottalization, and flapping all played a role. These results provide a characterization of what subphonemic components must be present for the categorical decision of /t d/ presence or absence that MFA makes.

The remainder of this introduction reviews forced alignment technology (subsection 4.1.1) and how this technology is useful to linguists (subsection 4.1.2) before examining the nature of /t d/ deletion in English (subsection 4.1.3). This section then concludes with the research questions arising from the background literature (subsection 4.1.4).

4.1.1 Forced Alignment

Speech technology can be divided into two components: speech synthesis (including text-to-speech) and speech recognition. While speech synthesis and text-to-speech involve "converting strings of text words into acoustic waveforms" (Jurafsky & Martin, 2019, ch. 27, p. 1), speech

of the signal at points in time; the audio signal is interpolated to be continuous but isn't *truly* continuous. This is all being said to acknowledge the shortcomings of categorical judgments, but also to say that at some point, changes in the fine phonetic detail will be reflected in a categorical measurement.

recognition is the reversal of this process: "transcribing acoustic waveforms into strings of text words" (Jurafsky & Martin, 2019, ch. 27, p. 1). An offshoot of speech recognition is forced alignment. Forced alignment is different from speech recognition because forced alignment also requires a transcript to be provided in addition to a sound file. The computer "aligns" the words it receives with the speech, while automatic speech recognition provides a transcript.

There are several different forced aligners: Prosody-Lab Aligner (Gorman, Howell, & Wagner, 2011), Penn Phonetics Forced Aligner (P2FA) (Yuan & Liberman, 2008), Forced Alignment and Vowel Extraction (FAVE) Align (Rosenfelder, Fruehwald, Evanini, & Yuan, 2011), SPeech Phonetization Alignment and Syllabification (SPPAS) (Bigi & Hirst, 2012), Munich AUtomatic Segmentation (MAUS) (Kisler et al., 2017), and Montreal Forced Aligner (MFA) (McAuliffe et al., 2017), among others. Prosody-Lab Aligner, P2FA, FAVE and MAUS differ from MFA in that the former are based on HTK ASR (Young et al., 2002) while MFA is based on Kaldi ASR (Povey et al., 2011). Gonzalez, Grama, and Travis (2020) found MFA to be more accurate than FAVE, MAUS, and Prosody-Lab Aligner, perhaps due to this difference. In working with the Digital Achieve of Southern Speech (DASS) (Kretzschmar et al., 2013; Kretzschmar et al., 2019), we have also found that MFA aligned DASS better than ProsodyLab.

In order to understand better how forced alignment works, the processes will be described in the subsequent paragraphs. The first step in forced alignment is feature extraction. Because the wave is continuous, it must be parameterized into discrete speech vectors. As such, forced alignment works by using signal processing techniques to extract features in frames of time from the wave. For example, MFA uses a 25 ms frame with 10 ms overlap. These features are most often Mel Frequency Cepstral Coefficients (MFCCs), but others, such as perceptual linear prediction (PLP) coefficients (Hermansky, 1990), could be used. MFA uses thirteen MFCCs, making for a total of 39 features including delta and delta-delta features (first and second derivatives, respectively, of MFCCs).

ASR systems typically use MFCCs in order to measure the physical properties of a speech wave; however, it is not transparent what these features *acoustically* correspond to, due to their calculation. Therefore, this section provides an overview of how MFCCs are calculated, in order to provide more transparency to the reader.

In the calculation of MFCCs, a speech wave is separated into shorter segments (assumed to be static), called frames, with overlap. Typically, and as is the case in MFA's default settings, these frames are 25 ms long with 10 ms overlap. These frames are then subjected to the Fourier transform, which computes base frequencies from a sound wave in order to produce a spectrum[2]. This step takes a waveform in the time domain (graphically represented with x = time and y = intensity) and converts it to a frequency domain (where instead, the x axis is *frequency*). These frequencies of the individual sine waves that together make up the complex wave of the speech sound are then represented as peaks in the spectrum[3].

This spectrum is then "mapped" onto Mel scale (Stevens, Volkmann, & Newman, 1937), which is a logarithmic scale that accounts for the perceptual differences in pitch[4], using triangular overlapping windows known as the Mel-filter bank, shown in Figure 4.1[5].

The result is a Mel spectrum, the coefficients of which are calculated by taking a weighted sum of the energy of the spectrum around each of the frequencies defined by the filterbank, and the log of these values is taken. Because the coefficients in the Mel spectrum are correlated, the discrete cosine transform (Ahmed, Natarajan, & Rao, 1974) is taken to decorrelate them, producing *cepstral*[6] coefficients (see Bogert (1963))[7]. These features (MFCCs) that were

[2]The Fourier Principal states that complex wave can be broken down into individual sine waves, which would have different frequencies based on wavelength ($f = v/\lambda$).

[3]A spectrum, rotated 90 degrees, makes a spectrogram, where higher amplitude peaks are shown as darker regions on the spectrogram.

[4]Pitch is expressed as frequency, but pitch is *logarithmically* dependent on frequency (Stevens & Volkmann, 1940).

[5]These triangular windows are narrower at lower frequencies and wider at higher frequencies, as the differences at higher frequencies are not as perceptually relevant.

[6]"Spec" backwards is "ceps".

[7]Appendix A in Rao and Manjunath (2017) contains the mathematical formulas for and more information about the calculation of MFCCs.

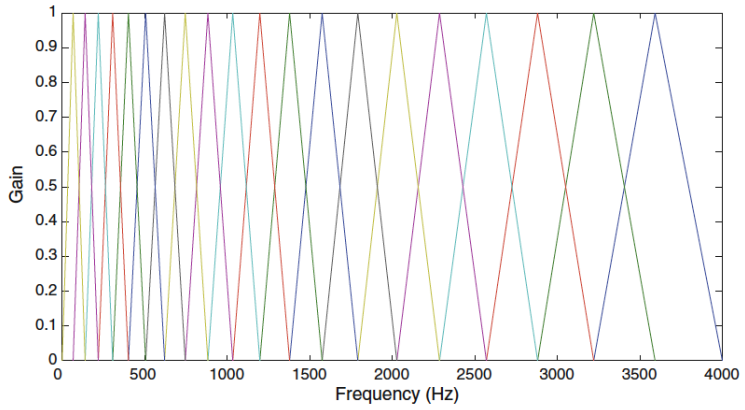

Figure 4.1: Mel-filter bank, reproduced from Rao and Manjunath (2017, p. 87)

extracted from the wave must then be mapped to phones in the acoustic model, in a phone recognition process. These acoustic models consist of observations from training data that is preferably hand-labeled, and these observations are modeled in Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs).

A Hidden Markov Model can model sequential data, in this case, acoustic features of each frame. It follows the Markov assumption that previous states do not have an effect on the transition out of the current state. An HMM consists of a set of states, Q (including two special states, a start and end state), a matrix that gives transition probabilities between states, A, the observations modeled O, and a set of emission probabilities. An example of an HMM is shown in Figure 4.2.

In modeling speech, a state could correspond to a phone (an example of which is shown in Figure 4.3, showing an HMM for the word *six*) or parts of a phone (an example of which is shown in Figure 4.4, showing an HMM with a beginning phone state, a middle (i.e., steady) phone state, and an end phone state), since a phone's acoustics are not static throughout the duration of the phone and consist of transitions and a steady state. A triphone model

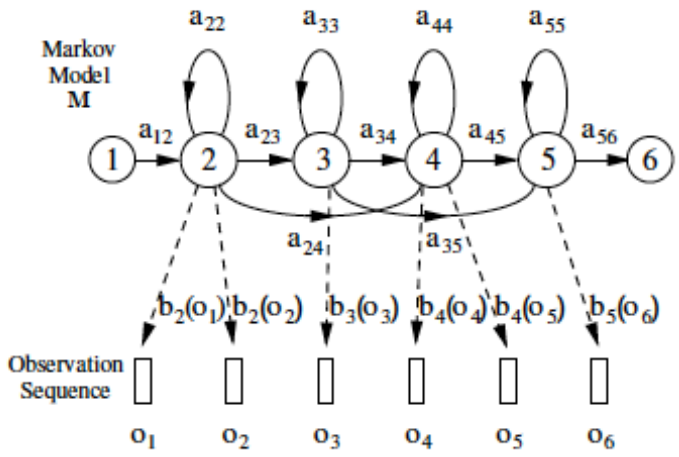

Fig. 1.3 The Markov Generation Model

Figure 4.2: Hidden Markov Model, reproduced from Young et al. (2002, p. 4)

uses HMMs that incorporate the beginning, steady, and end state of a phone, and MFA uses a triphone model in order to account for acoustic context on either side of a phone.

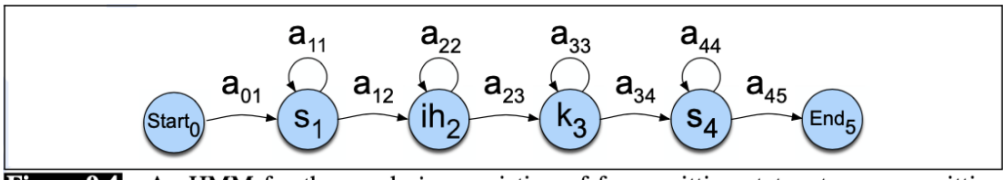

Figure 9.4 An HMM for the word *six*, consisting of four emitting states, two non-emitting states, and the transition probabilities A. The the observation probabilities *B* are not shown.

Figure 4.3: HMM for the word *six*, reproduced from Jurafsky and Martin (2008, p. 294)

Within each state q, there are GMMs that models probability of values of a feature vector given a phone, $\Pr(o \mid q)$ [8]. This GMM consists of multiple Gaussian distributions, each with their own weight, (i.e., a *mixture* of Gaussian distributions) for each of the 39 features.

In automatic speech recognition, output is estimated by calculating prior probability based on a language model (i.e., probability of phrases) and observation probability, based on

[8] Equivalent to the probability of a phone given a feature value $\Pr(q \mid o)$, due to Bayes Theorem, $\Pr(o \mid q) = \frac{\Pr(q|o)}{\Pr(o)\Pr(q)}$.

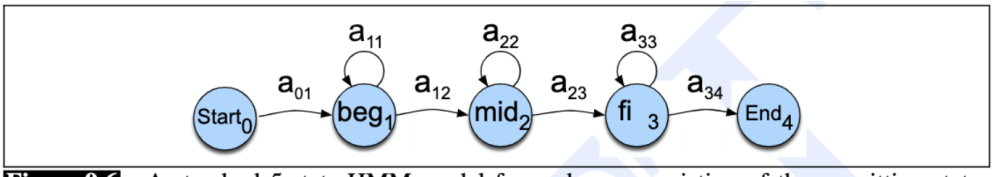

Figure 9.6 A standard 5-state HMM model for a phone, consisting of three emitting states (corresponding to the transition-in, steady state, and transition-out regions of the phone) and two non-emitting states.

Figure 4.4: HMM with transition and steady states, reproduced from Jurafsky and Martin (2008, p. 296)

the acoustic model. In the case of forced alignment, we do not use a language model because we already know the phrases present in the data from the transcription required. Instead, we have a lexicon (dictionary) consisting of words and the phones that comprise them. These dictionary entries are represented in the same way that HMMs are, as finite-state transducers (FSTs). A finite-state transducer is a finite-state machine, which is a model that has "states" and transitions between these states. When given an input, it produces an output and can be thought of as a type of translator. In the dictionary, the finite-state transducer maps input to output; in this case, mapping words to their pronunciation. An example of the dictionary pronunciation given for *just* is given in Figure 4.5. The pronunciation for *just* is given as either "J AH1 S T" or "J IH0 S T". The acoustic model, combined with word pronunciations, is decoded with the Viterbi algorithm (Viterbi, 1967) to obtain the most probable alignment given the acoustics and phones.

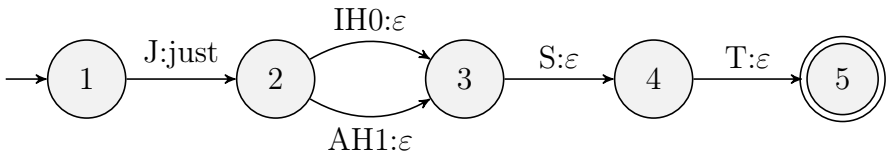

Figure 4.5: Example FST of *just*

4.1.2 Application of Forced Alignment for Linguists

While forced alignment and automatic speech recognition is typically judged on word error rate (Jurafsky & Martin, 2000, p. 328), the boundaries of individual segments are of great importance to linguists, where correct boundaries matter in measuring acoustic variables of segments, such as voice onset time. One problem that has been recognized in forced alignment is that it is indeed *forced*. The system will provide an alignment for phones, as listed in the dictionary, that are not present in the acoustics (Dall et al., 2016). It follows, then, that a dictionary based on canonical pronunciations found in General American English may not capture the phonetic realization present. Previous studies have shown that modification of forced alignment has produced more accurate forced alignment results. For example, Yuan and Liberman (2011) found that by training two acoustic models, one featuring g-dropping and one without g-dropping, the aligner's agreement with human transcribers was near the rate of intertranscriber agreement. Schuppler (2011) found that modification of the dictionary HTK (Young et al., 2002) was able to improve forced alignment of conversational Dutch so long as dictionary modifications were made with sensitivity to the conditions that made reductions likely to occur. Additionally, the AudioBNC has a custom dictionary designed to maximize dialect coverage (Coleman, Baghai-Ravary, Pybus, & Grau, 2012). Bailey (2016) found that modification of the dictionary used in the processing of sound files by FAVE (Rosenfelder et al., 2011) was able to detect h-dropping, /t d/ deletion, and th-fronting, and he found that performance was affected by voicing of the segment that underwent the phonological process and also by speech rate. Additionally, Shi (2019) were able to improve the results of MFA (McAuliffe et al., 2017) on the Digital Archive of Southern Speech (Kretzschmar et al., 2013; Kretzschmar et al., 2019) by modifying the dictionary to better align the *pin-pen* merger that occurs in Southern American English.

As the current study is similar to Bailey's (2016) study, I will outline crucial differences here. First, we use different corpora, as I use the Buckeye Corpus (consisting of Ohio English),

while Bailey uses a small Manchester English corpus; the different corpora feature different varieties of English: American and British English. /t d/ deletion in these different varieties is thought to be conditioned by different factors (Tagliamonte & Temple, 2005). Furthermore, we use different aligners: he uses FAVE while I use MFA. These aligners have different underlying technology, as FAVE uses HTK and MFA uses Kaldi.

4.1.3 Word-final /t d/ Deletion

Word-final /t d/ deletion has been studied in many varieties of English, e.g., African American English (Labov, 1968; Wolfram, 1969), British English (Tagliamonte & Temple, 2005; Tanner, Sonderegger, & Wagner, 2017; Temple, 2014), Chicano English (Santa Ana, 1991), Tejano English (Bayley, 1994), and General American English (Guy, 1980; Tamminga, 2018); it is a well-attested sociolinguistic variable. Additionally, numerous linguistic factors have been said to influence /t d/ deletion: morphological status, following segment, preceding segment, stress, frequency, phonological neighborhood density, and voicing. The sections that follow detail these aforementioned linguistic factors.

4.1.3.1 Morphological Status

Cohen and Labov (1963), Wolfram (1969) and Guy (1980) all found distinctions in /t d/ deletion based on morphological properties (i.e., monomorphemic words vs. bimorphemeic words). Monomorphemic words were more likely to feature /t d/ deletion than semi-weak verbs (verbs that are weak but also featured a vowel change, e.g., *sleep* $\sim$ *slept*, than regular past tense verbs, e.g., *walk* $\sim$ *walked*). However, Tagliamonte and Temple (2005) did not find an effect of morphological status in British English /t d/ deletion, which the authors attribute to language-specific variation and *not* a phonetic process. They claim this deletion occurs due to the following and preceding segments, discussed in the following two sections. A very recent study done on the Librispeech Corpus found that rate of /t d/ deletion in

each morphological class, calculated in a token-based[9] way, produced similar results to the aforementioned literature that claim deletion rates highest for monomorphemes, followed by semi-weak past tense, followed by regular past tense (Yuan et al., 2020).

4.1.3.2 Following Segment

For following segment, obstruents are expected to have highest rates of deletion, followed by liquids, followed by glides, followed by vowels, followed by pauses. Similarly, Santa Ana (1991) also found an inverse relationship between sonority of following segment and deletion rates in Chicano English, and Raymond, Brown, and Healy (2016) found word-final /t d/ to be more likely to be deleted if it preceded another consonant. Again, Yuan et al. (2020) produced similar findings to Guy (1980), Santa Ana (1991) and Raymond et al. (2016): /t d/ deletion rates were higher when the /t d/ preceded consonants, followed by glides and liquids, and vowels and pauses had lowest rates of deletion, as shown in Figure 4.6.

4.1.3.3 Preceding Segment

A third factor predicting /t d/ deletion is the preceding segment, with deletion being most prevalent after /s/[10], then stops, nasals, non-sibilant fricatives, and laterals. However, deletion did not occur after /ɹ/ (Guy, 1980). Santa Ana (1991) found a similar pattern to hold true in Chicano English as well, with less sonorous preceding sounds to feature more deletion in word-final /t d/. As following segment and preceding segment patterns hold for so many varieties, Tagliamonte and Temple (2005) argue that is a universal phonological phenomenon. Yuan et al. (2020) found that /t d/ following nasals or coronal obstruents were more likely

[9]In the Yuan et al. (2020) study, /t d/ deletion in morphological class was calculated in two ways: token-based and type-based. In the token-based way, the rate of deletion was calculated "on all word tokens in the same class", while in the type-based way, they "calculated a deletion rate for every word type, and then calculated the means of the rates of all word types" (2020, p. 7326).

[10]Although it was not included in the main part of his analysis, Guy (1980) found that as the number of changes in articulatory position increase, the rates of deletion increased (e.g., /kst/ features one change, /skt/ features two). Furthermore, Guy (1980) suggests that a longer consonant cluster is more likely to feature deletion (i.e., *mixed* is more likely to feature deletion than *mist*). However, this was not included in the main part of his study either.

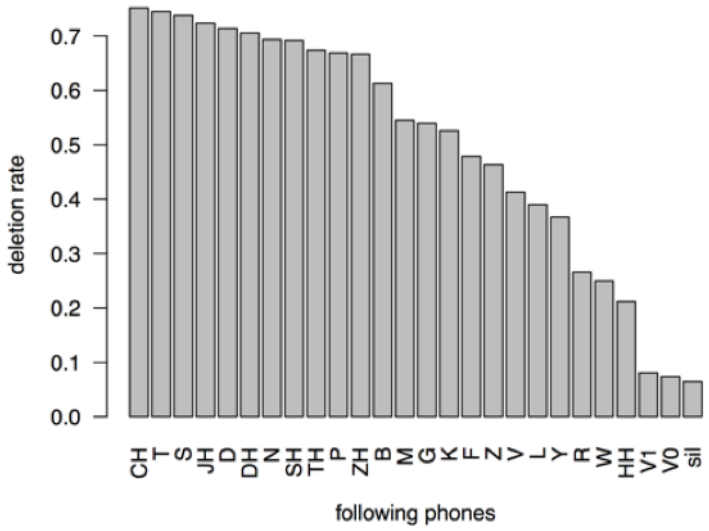

Figure 5: *Effect of the following phone.*

Figure 4.6: Deletion rate based on following phone, reproduced from Yuan, Lin, and Liu (2020, p. 7326)

to be deleted, and similarly to the aforementioned studies, found high rates for /s/ and low rates for /ɪ/, as shown in Figure 4.7.

4.1.3.4 Stress

Fasold (1972) found /t d/ deletion more likely to occur in unstressed syllables, in line with research that shows stressed segments are less likely to be reduced and therefore more likely to be a full, more canonical form either temporally (Fry, 1955) or spectrally (Jones, 1956)[11].

4.1.3.5 Frequency

Frequency has been established to play a role in reduction (Bybee, 2002; Jurafsky & Martin, 2000). Specific to /t d/ deletion, Raymond et al. (2016) did not find frequency to have

[11]See Lindblom (1963) for an overview.

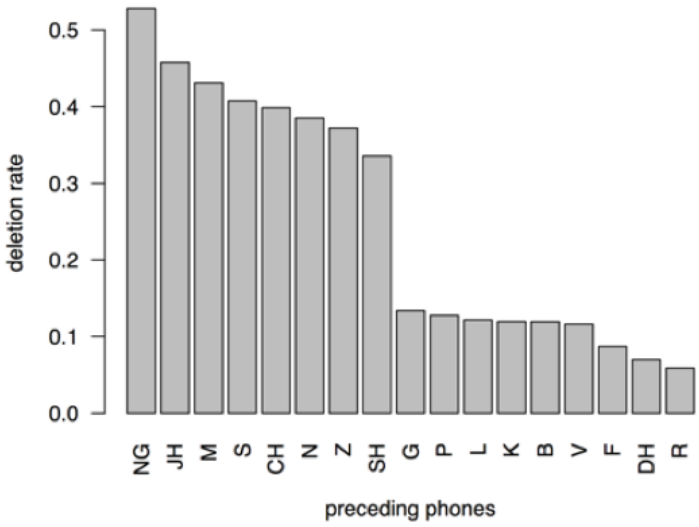

Figure 4: *Effect of the preceding phone.*

Figure 4.7: Deletion rate based on previous phone, reproduced from Yuan, Lin, and Liu (2020, p. 7326)

an effect on /t d/ deletion. However, this finding is in contradiction with Guy, Hay, and Walker (2008), who found frequency effects influenced /t d/ deletion in New Zealand English; Renwick, Baghai-Ravary, Temple, and Coleman (2014), who found lexical frequency to play a role in British English; and Yuan et al. (2020), who found higher frequency words (as measured by base10 logarithm of word frequency) to have more /t d/ deletion. However, in Yuan et al.'s (2020) mixed-model analysis, frequency was only significant in an interaction with morphological status.

4.1.3.6 Phonological Neighborhood Density

In a study of word duration and vowel centralization (i.e., reduction) in the Buckeye Corpus (Kiesling et al., 2006), Gahl et al. (2012) found that phonological neighborhood density increased likelihood of reduction. Phonological neighborhood density is a measure of how many words in the lexicon are similar to a given word (e.g., off by one segment, including

deletion of segments. For example, the phonological neighbors of *cat* are *chat*, *cap*, *at*, etc. Yuan et al. (2020) found words with higher phonological neighbor density were more likely to exhibit /t d/ deletion.

4.1.3.7 Voicing

Yuan et al. (2020) found that the voiced counterpart of /t d/ deletion is more likely to be deleted than the unvoiced. In fact, Bailey (2016) found lower rates of accuracy in forced alignment judging a /t/ to be present when it was present, as shown in Figure 4.8.

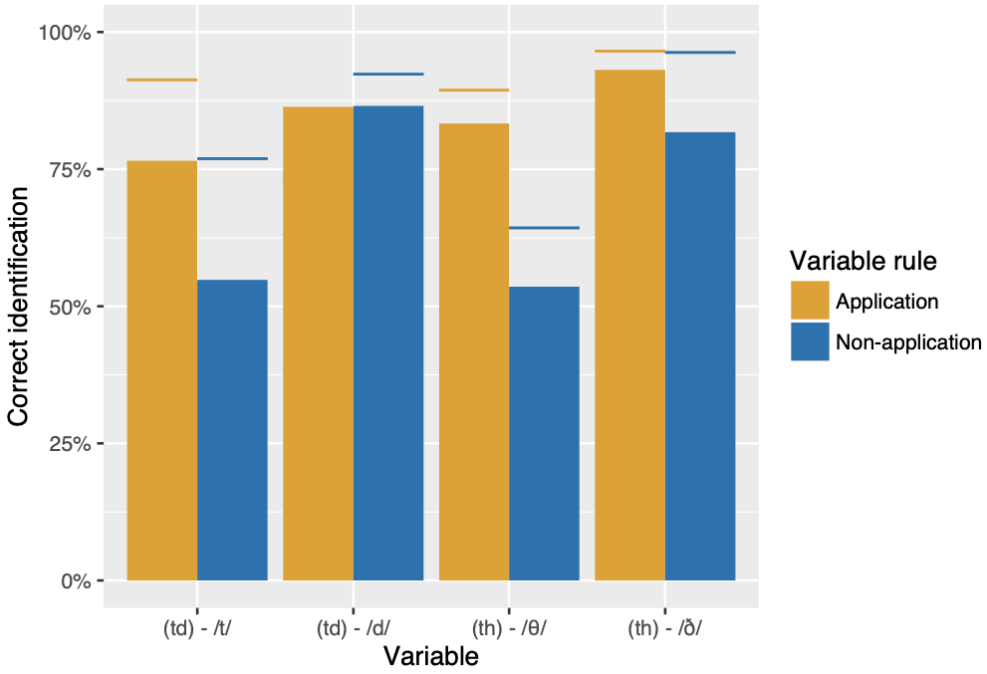

Figure 4.8: FAVE accuracy rates for voiced and voiceless segments by rule application, reproduced from Bailey (2016, p. 16)

4.1.3.8 In the Buckeye Corpus

Because the Buckeye Corpus is used to investigate MFA's performance on /t d/ deletion, this section discusses previous research on the phenomenon in the corpus. There have been studies dealing with rates of word-internal /t d/ deletion (Raymond, Dautricourt, & Hume, 2006),

which also gives rates of /t d/ deletion in the Buckeye Corpus (Kiesling et al., 2006) as 16.5% (and non-canonical realizations as 45%). Additionally, word-final /t d/ in monomorphemes was found to be deleted 53% of the time, in weak verbs 41% of the time, and in regular past tense 23% of the time (Tamminga & Fruehwald, 2013).

4.1.4 Research Questions

Previous research suggests HTK-based aligners work better when modifications to the dictionary are based on phonological variation. My research questions are as follows: will MFA's acoustic models detect /t d/ deletion rates similar to human-annotated data and/or inter-transcriber agreement in speech from the Buckeye Corpus (Kiesling et al., 2006) if the dictionary is modified to allow for word-final /t d/ deletion? Furthermore, where is MFA likely to fail, and what governs this inability? My hypothesis is that MFA will be able to detect word-final /t d/ deletion with dictionary modification at a rate between human ability and inter-transcriber agreement, and that variables affecting likelihood of /t d/ deletion (as previously discussed, morphological status, following segment, preceding segment, stress, frequency, phonological neighborhood density, and voicing of /t d/) may affect MFA's accuracy in determining /t d/ deletion. Additionally, I predict that the areas of lower accuracy rates will have phonetically/acoustically-grounded explanations. An additional factor to consider is speech rate. Bailey (2016) found that FAVE's accuracy decreased as speech rate[12] rose, unlike the rate of human transcribers (which remained constant), so this effect will be examined as well. An additional research question that this chapter answers is, what subphonemic properties of /t d/ are likely to influence MFA's judgment of [t, d] presence? My hypothesis is that the burst will play a central role in MFA's judgment that a stop is present, due to

[12]Speech rate does not necessarily correspond to speech style, which is thought to affect reduction (as discussed in Chapter 1). More casual speech is more likely to feature reduction (Ernestus, 2000), but casual speech is not necessarily slower or faster than more formal speech. In this study, all speech is considered semi-casual interview speech.

the abrupt change in the acoustics from the closure (or preceding segment, if a closure is not present).

4.2 Methodology

This section will first discuss MFA and the inputs it requires (subsection 4.2.1), followed by an overview of the Buckeye Corpus (Kiesling et al., 2006) (subsection 4.2.2). Furthermore, selection of data (subsection 4.2.3), variables (subsection 4.2.4), and subphonemic analysis (subsection 4.2.5) are described in this section.

4.2.1 MFA

As previously mentioned, the Montreal Forced Aligner (MFA) was used as Gonzalez et al. (2020) found it to be more accurate than FAVE (Rosenfelder et al., 2011), MAUS (Kisler et al., 2017), and Prosody-Lab Aligner (Gorman et al., 2011). MFA requires a dictionary of words and pronunciations consisting of phones, an acoustic model, wav files, and their corresponding transcripts.

The Librispeech dictionary (Panayotov, Chen, Povey, & Khudanpur, 2015) was used as suggested in MFA documentation[13]. The dictionary, containing 206,508 words, was modified to add /t d/ deletion for words where it was most likely to occur: in words ending with a /t/ or /d/ that was preceded by a consonant. This resulted in an addition of 14,754 words, for a total of 221,262 words. These dictionary entries are modeled as finite state transducers (FSTs). For example, in the Librispeech dictionary, there are two entries for *just*, given as "J AH1 S T" and "J IH0 S T". In the new dictionary, there are four entries for *just*: "J AH1 S T", "J IH0 S T", "J AH1 S", and "J IH0 S". An FST for the new dictionary is shown in Figure 4.9 which can be compared to an example of *just* in an unmodified dictionary, given

[13]MFA's documentation does not explicitly say why they suggest Librispeech, though I believe it has something to do with greater coverage. The Librispeech dictionary has 206,508 words, while the CMU dictionary only has 133,801 words.

in Figure 4.5. In these figures, an end state is indicated with two circles. Following Bailey (2016), the aligner was not retrained after dictionary modification[14].

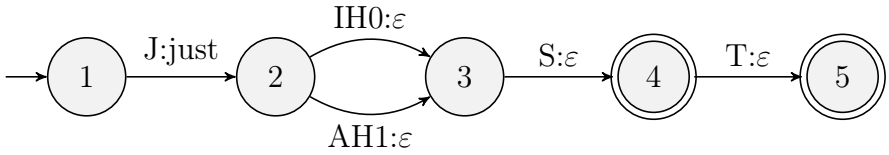

Figure 4.9: FST in modified dictionary for *just*

The pre-trained acoustic model (trained from Librispeech corpus data (Panayotov et al., 2015)) was used. These acoustic models were trained using MFA's standard training procedure, shown in Figure 4.10, which uses triphone models.

Next, the Buckeye Corpus wav files were used, but because the Buckeye Corpus transcriptions are given in ESPS *Aligner* format, these were converted to a Praat (Boersma & Weenink, 2018) TextGrid format[15] by a Python script for use by MFA.

4.2.2 Buckeye Corpus

The Buckeye Corpus (Kiesling et al., 2006) consists of approximately 300,000 words of conversational speech of 40 native central Ohio speakers. These speakers are balanced for age of speaker (over 40, under 40), gender of speaker (female, male), and gender of interviewer (female, male). The Buckeye Corpus is used due to its transcription procedure. The alignment was produced by the ESPS *Aligner* software, but adjustment of the alignment was made by phonetically-trained human labelers. According to the manual for the Buckeye Corpus, these humans went through the following process:

[14]In Shi (2019), a study that modified the dictionary of MFA detect a prenasal merger, the acoustic model was retrained after dictionary modification. This was done in order to update acoustic models to reflect the fact that there was a "new" phone — IH that preceded NG (as opposed to "regular" IH). In this case, there is no new phone, though including other variants of /t/ and retraining acoustic models may improve results.

[15]MFA also accepts plain text files, but TextGrids were used in order to time align utterances for improved accuracy.

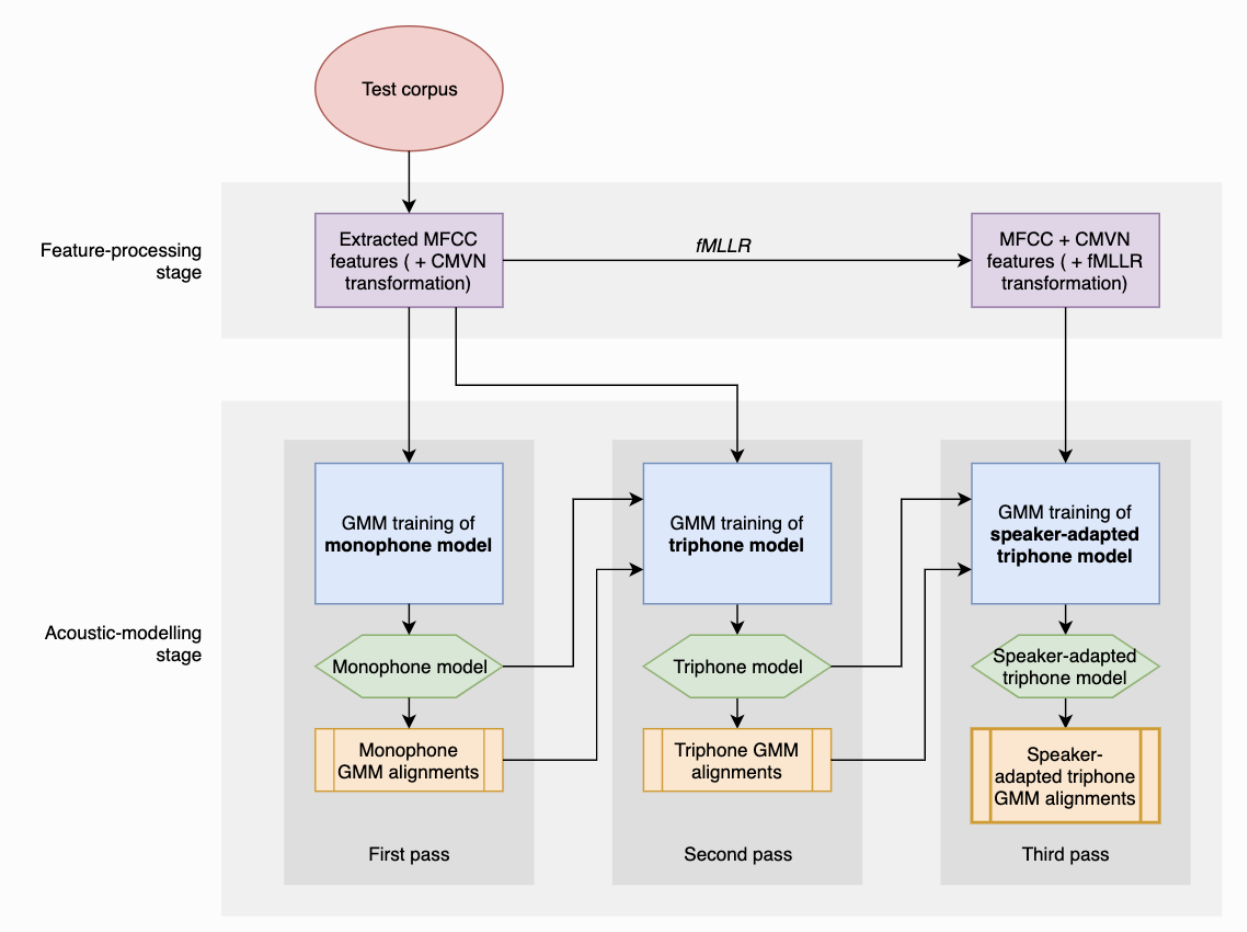

Figure 4.10: MFA training process, reproduced from the documentation

Manipulating the temporal position of labels, as well as adding, deleting or changing labels to be consistent with the evaluation of a phonetically trained human labeler. Human labelers used speech spectrogram and waveform displays generated using Xwaves (formerly from Entropics Inc.) software as well as auditory perceptual evaluation in determining phone labels (Kiesling et al., 2006).

The alignment produced by human labelers can then be considered a gold standard for comparison to the results of MFA[16]. Additionally, the Buckeye Corpus is used because it contains natural speech, which typically exhibits more variation and reduction (Ernestus & Warner, 2011; Keating, 1998).

4.2.3 Data Selection

R (R Core Team, 2018), RStudio (RStudio, 2017), and the R packages 'tidyverse' (Wickham, 2017) and 'fuzzyjoin' (Robinson, 2019) were used to process the data. The first step was to identify the tokens that may be likely to undergo word-final /t d/ deletion. Using the transcription data from the Buckeye corpus, a Python script was used to find words whose canonical pronunciations end with a consonant followed by a /t d/ and the start and end times for these words, and 24,219 tokens were found. Because the start and end times of a human aligner and MFA wouldn't be expected to line up exactly, there was a 100 ms allowance on each side of the human-aligned end time given to match the word with MFA results. Additionally, there was the potential for MFA to be unable to align a segment (marked by <unk>). A total of 23,522 tokens remained after filtering. There was some additional data not present in this analysis. One file, 1901b had a wav file that was significantly shorter than the alignment provided by the Buckeye Corpus, and so this file was excluded. One Buckeye alignment, 2801a, had time stamps that overlapped (i.e., one utterance was given an end timestamp of 431.847408, while the next period started at 421.847408) and this file was excluded as well.

4.2.4 Variables

The following section outlines the variables measured that may influence /t d/ deletion.

[16]However, Milne points out that "[f]orced alignment should not seek to match human performance, because human performance is variable in ways that forced alignment is not and should not be (i.e., in the way Mitterer (2011) details" (2014, p. 5).

4.2.4.1 Morphological Status

To code words for morphological status, stemming was used. Stemming is the process of stripping affixes from a word, resulting in the stem. The R package 'SnowballC' was used to implement the Porter Stemmer (Porter et al., 1980) in R to determine whether the word was a monomorpheme. Additionally, the Buckeye Corpus is part-of-speech tagged using the Penn Treebank tag set (Marcus, Santorini, & Marcinkiewicz, 1993). If the part of speech was tagged as 'VBD' (verb, past tense), 'VBG' (verb, gerund/present participle), or 'VBN' (verb, past participle) and the orthographic word ended in <-ed>, it was considered a regular verb. If the part of speech was tagged as 'VBD', 'VBG', or 'VBN' and the orthographic word did not end in <-ed>, it was considered an irregular verb. Then, if the Porter Stemmer did not stem the word, it was coded as a monomorpheme, and otherwise, it was classified as "other".

4.2.4.2 Preceding and Following Segment

Since we are looking at reduction in consonant clusters, the canonical phonemic pronunciation given by the Buckeye Corpus was used to determine preceding phone. However, actual pronunciation given by the Buckeye Corpus was used to determine following phone, especially so that pauses could be recorded in this variable.

4.2.4.3 Stress

The pronunciations given by the Librispeech dictionary feature stress, so the pronunciation returned by MFA was used to determine the stress of the last syllable of the word.

4.2.4.4 Frequency

To calculate frequency, the Python package 'wordfreq' (Speer, Chin, Lin, Jewett, & Nathan, 2018) was used to obtain Zipf frequency, the "base-10 logarithm of the number of times it appears per billion words" (Speer et al., 2018), proposed by Brysbaert and New (2009,

n. p.). For English, the measure of frequency is based on Wikipedia, subtitles (from OPUS OpenSubtitles 2018 (Lison & J., 2016) and SUBTLEXus), news (from NewsCrawl 2014 (Barrault et al., 2020)[17] and GlobalVoices (Tiedemann, 2012)), books (from Google Books Ngrams 2012 (Michel et al., 2010)), web text (from ParaCrawl (Esplà-Gomis, Forcada, Ramírez-Sánchez, & Hoang, 2019) and the Leeds Internet Corpus (Sharoff, 2016)), Twitter, and Reddit.

4.2.4.5 Phonological Neighborhood Density

To calculate phonological neighborhood density, canonical pronunciations from the Buckeye Corpus were converted to one-letter ARPABET characters, as were the dictionary entries from the Librispeech corpus. The R package 'stringdist' (van der Loo, 2014) was used to calculate the Levenshtein distance between each word and each entry in the Librispeech dictionary. Then, the sum of words exhibiting a distance of 1 was base-10 log transformed.

4.2.4.6 Voicing

I list this for completeness: the last variable, voicing of /t d/, was classified as voiced for [d] and voiceless for [t].

4.2.4.7 /t d/ Deletion Measurement

In the Buckeye Corpus, /t d/ deletion was considered to occur if any other segment besides /t d/ or one of six variants, /ð θ ɾ t̪ d̪ tʃ dʒ/, was present word-finally in the pronunciation given by transcribers of the Buckeye Corpus. There were 220 instances of devoicing (/d/ → [t]) and 114 instances of voicing (/t/ → [d]), which were also categorized as not featuring /t d/ deletion.

[17]Released as part of EMNLP 2020 Fifth Conference on Machine Translation (WMT20).

These common pronunciation variants and their counts are listed in Table 4.1[18].

Table 4.1: Pronunciation variants

Variant	count
ð	10
θ	3
ɾ	492
t̰ or d̰[19]	1553
tʃ	28
dʒ	10

4.2.4.8 Speech Rate Measurement

Speech rate was determined by calculating the average phones per second for each speaker.

4.2.5 Subphonemic Detail

In order to better understand MFA's performance and its relationship to the fine phonetic detail of the acoustics, error analysis of a randomly selected 10% of each of four subsets of tokens was done. The number of tokens in each group are as follows: 484 /t/s that MFA categorized as present, 665 /t/s that MFA categorized as absent, 477 /d/s that MFA categorized as present, and 726 /d/s that MFA categorized as absent) were examined by hand. The acoustics of each token were examined and judged on constriction (present or absent), constriction frication (frication, no frication, not applicable), constriction voicing (voiced, partially voiced, voiceless, not applicable), burst (present or absent), number of bursts (one, multiple, not applicable), and strength of burst (weak, strong, not applicable), similarly to Schuppler (2012)[20].

[18]The manual for the Buckeye Corpus includes instructions on the transcription of glottalization of /t d/ as 'tq', the palatalization of /t d/ (usually after /r/) as 'ch' and 'jh' respectively, and the oral flaps or taps as 'dx' .

[19]Seyfarth and Garellek (2020) found singleton coda /t/ glottalization in the Buckeye Corpus to occur before almost all sonorant onsets (with the exception of /j/).

[20]However, Schuppler (2012) developed additional criteria based on the context of /t/ that were not used in this study.

Determination of the presence of a constriction was based on a decrease in amplitude corresponding with the end of the previous segment. The constriction was determined to end when there was a significant change in the spectrogram or waveform that corresponded to either a burst/delayed release or the following segment. The constriction was classified as either voiced or voiceless, based on the presence of a voicing bar in the spectrogram. Figure 4.11 shows an example of a /d/ with a clear closure, and partial voicing (bleed from the preceding segment).

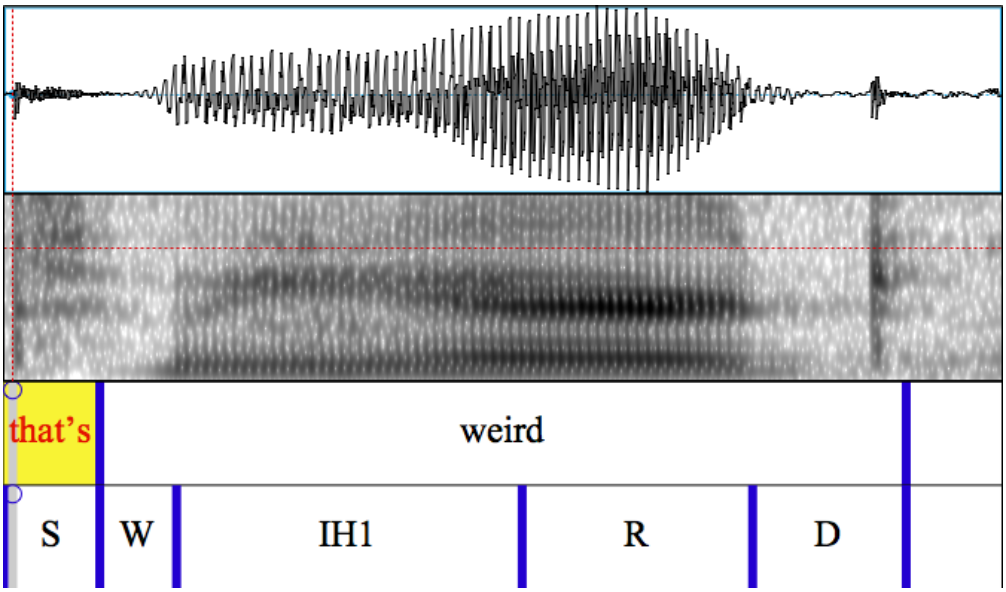

Figure 4.11: Partially voiced closure of /d/

The constriction was also classified as either having frication or not having frication based on the waveform's periodicity. If there was aperiodic noise in the closure, it was classified as having frication[21]. An example of frication during the closure of /d/ is shown in Figure 4.12.

The burst was classified as present or absent primarily based on the presence of a vertical band of energy on the spectrogram, similarly to Schuppler (2012). The waveform and auditory

[21]A periodic wave is one that repeats its period (the time in which one cycle of the pressure variation of a wave is complete) and occurs in vowels, liquids, glides, and nasals. An aperiodic wave is a wave in which this cycle does not repeat, and this kind of frication noise is the result of a partial obstruction in the vocal tract. Note that the signal can be "mixed-source", being both aperiodic as a result of turbulent airflow in the vocal tract and periodic as a result of the vibration of vocal folds, as is the case with voiced fricatives (e.g., /z/).

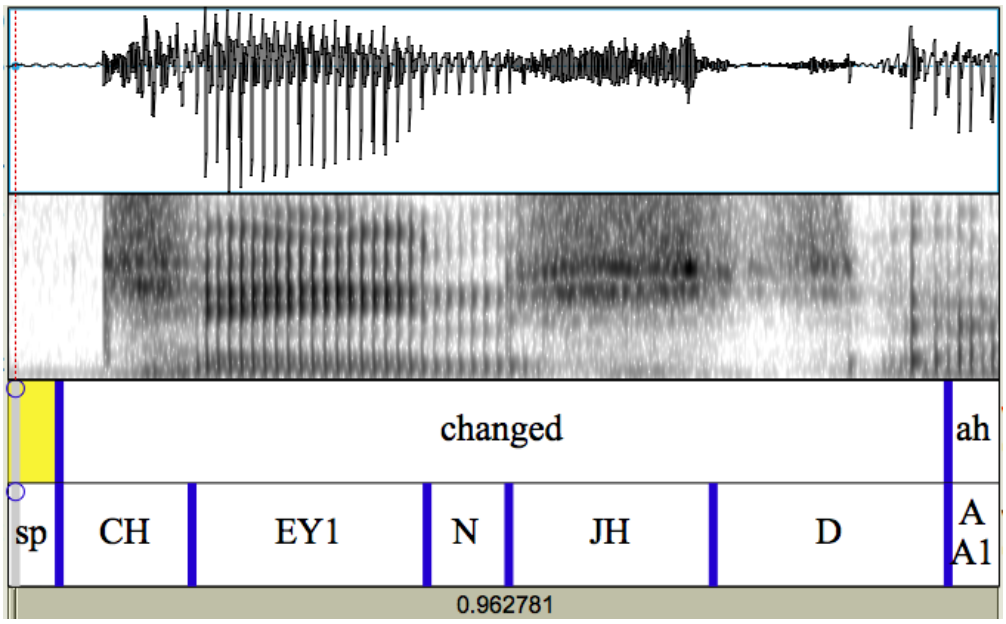

Figure 4.12: Frication during closure of /d/[a]

[a]It is also worth noting that the boundary of the /d/ is too far to the right, and it includes the glottal onset of the following vowel.

qualities were also examined, though as Olive, Greenwood, and Coleman (1993) point out, bursts are not always apparent in the waveform, especially if they are weak. The stop was considered to have multiple bursts if there were two distinct bursts with both a vertical band on the spectrogram. Next, bursts were classified as weak if they only had energy in part of the spectrogram and/or had a very short duration, or they were classified as strong if energy was present at all frequencies, again, based on Schuppler (2012). In the event of multiple bursts, the strength of the first burst was recorded. Figure 4.12, in addition to showing frication during the closure, also shows a weak burst, as the energy does not form a complete band in the spectrogram. Figure 4.11 exhibits a strong burst; energy is present in a vertical band in the spectrogram. Additionally, the presence of a delayed release was determined by the existence of aperiodic noise after the burst or closure.

Instances where a nasal preceded the /t/ or /d/, a fricative followed, and featured deletion of the /t/ or /d/ (e.g., *and so*) would then be the kind of context in English that gives rise to excrescent stops[22]. These instances were classified as excrescent bursts. Due to the nature of a sequence of [n] + [d], there is not a closure present for [d] (only for [n]), as shown in Figure 4.13.

In cases where a word-final stop was followed by a word-initial stop, (e.g., *and told*), the word-final /t/ or /d/ was only judged to be present if there was a burst between closures. In reality, the /t/ or /d/ could potentially be considered present based on the timing (i.e., if there was a longer closure period). This is a limitation of the current study.

Additionally, for the subsets of data where /t/ or /d/ was classified as absent, the presence of the closure, burst, and their respective components that were outlined above, were looked for by identifying the end of the segment preceding the word-final /t d/ (e.g., [n] of *and*) and the beginning of the next word, and looking between these two points in time.

While the above information refers to classification of /t/ and /d/, the information below is specific to /t/, as the pronunciation of word-final /t/ is expected to be highly variable (Olive et al., 1993). In addition to the categorization as discussed above, there were three additional options for /t/: glottalized, delayed release after /s/, and flapped.

A /t/ was marked as glottalized (and not containing any other designation as discussed above, i.e., closure, burst, etc.) if there was a somewhat abrupt stop to the preceding consonant followed by an increase in amplitude of the waveform and a vertical bar in the spectrogram that resembled a glottal pulse. An example of a glottalized /t/ is shown in Figure 4.14.

A /s/ + /t/ combination is one of the highly variable realizations Olive et al. (1993) allude to. Olive et al. (1993) go on to say the following regarding the waveform of /st/:

> In the waveform, the closure interval is not visible at all. In the lower plot [in Figure 4.15], the end of the fricative /s/ is signaled by a decay of the vibrations,

[22]Excrescent stops in English can occur between a nasal consonant and a voiceless fricative and could give rise to variants such as *warmth* /wɔɹm/ + /θ/ → [wɔɹmpθ].

> and the beginning of the stop /t/ is indicated by a slight rise in energy. However, because of the absence of a clear closure region, the sound does not really look like a stop. When we listen to the stop region in isolation (between 70 ms and 120 ms), we hear a sound that resembles the fricative /s/ more than the stop /t/. Yet in the context of the entire signal (between 25 ms and 120 ms), a two-phoneme cluster can definitely be heard. The second segment cannot be identified as a /t/ in isolation, and it is only contextual and lexical information that leads to the perception of a stop. Unlike the examples of word-initial which showed fully articulated stops, with clear closure regions and unaspirated bursts, these examples of word-final clusters are weakly articulated and variable (1993, p. 259–260).

As such, these instances were treated as a separate realization of /t/. Tokens of /t/ following /s/ that were perceived as /s/ + /t/ were recorded as exhibiting this phenomenon. In the event of an ambiguous token, the waveform and spectrogram were examine to see if there was a change in energy that would disambiguate the token.

Tokens of /t/ that were in the correct context to be flapped (e.g., *sort of*) were categorized as flaps based on the auditory presence of a flap combined with decrease in amplitude of the waveform and energy in the spectrogram.

The methodology of this study can be summarized as follows. MFA's dictionary was modified to accept variants of words that were likely to feature /t d/ deletion. Then, MFA was run on the Buckeye Corpus (which has human-checked transcriptions). Variables that influence /t d/ deletion (morphological status, preceding and following segment, stress, word frequency, phonological neighborhood density, and voicing) were recorded in order to see what effect (if any) they have on MFA. Further, fine phonetic detail in 10% of the corpus was manually examined to see what effect the presence of closure, burst, delayed release,

excrescent stops, flaps (for /t/), and glottalization (for /t/) had on MFA's judgment. The following section will give the results of this methodology.

4.3 Results

Looking at the Buckeye Corpus's human-corrected "gold standard" annotations alone, out of the 23,522 tokens selected for analysis, human-labeled annotations show that 16,796 featured /t d/ deletion and 6,726 had no /t d/ deletion, a /t d/ deletion rate of 71.4%. Because one token can vary in how it was transcribed by a human (absent vs. present) and how it was transcribed by MFA (absent vs. present), it is reasonable to look at the results in terms of true positive, false positive, true negative, and false negative. The first part of this term refers to the correctness of MFA; a true result is one where MFA is "correct", i.e., in agreement with the "gold standard" of the human transcriber, while a false result is one where MFA is "incorrect", i.e., not in agreement with the "gold standard" of the human transcriber. The second part of this term refers to whether the result was positive (in this case, the presence of /t d/ according to the "gold standard" human transcriber) or negative (the absence of /t d/ according to the "gold standard" human transcriber). Thus, a true positive is an instance in which MFA correctly chose a positive result, i.e., the presence of /t d/. A true negative is an instance in which MFA correctly chose a negative result, i.e., the absence of /t d/. A false positive is an instance in which MFA incorrectly chose a positive result, and a false negative is an instance in which MFA incorrectly chose a negative result. This information is summarized in Table 4.2.

The results, in terms of true positive, true negative, false positive, and false negative are given in Table 4.3. An additional table, Table 4.4, shows these counts as percents. The results show 5808 true positives (where MFA marked /t d/ present and the human marked /t d/ present), 10,894 true negatives (where MFA marked /t d/ absent and the human transcriber marked the /t d/ absent), 3802 false positives (where MFA marked /t d/ present and the

Table 4.2: Results and interpretation

Result	**Interpretation**		**Match/Mismatch**
	MFA	**Human**	
True positive	Present	Present	Match
True negative	Absent	Absent	Match
False positive	Present	Absent	Mismatch
False negative	Absent	Present	Mismatch

human transcriber marked /t d/ absent), and 3018 false negatives (where MFA marked /t d/ absent and the human transcriber marked /t d/ as present). Overall, MFA has an accuracy of 71%. Of course, because /t d/ deletion is not equally likely in all categories, MFA could perform better because of a difference in distribution. For this reason, Cohen's kappa (Cohen, 1960) was calculated, as this takes into consideration performance by chance, as discussed in subsubsection 2.2.6.1. A Cohen's kappa value of 0 would indicate that performance is as good as chance. For this data, $\kappa = 0.392$, considered by Landis and Koch (1977) to be "fair".

Table 4.3: MFA vs. human transcriber

		Human	
		∅	[t, d]
MFA	∅	10,894 (TN)	3018 (FN)
	[t, d]	3802 (FP)	5808 (TP)

Table 4.4: MFA vs. human transcriber, as percents

		Human	
		∅	[t, d]
MFA	∅	46.3% (TN)	12.8% (FN)
	[t, d]	16.2% (FP)	24.7% (TP)

The following results are separated by variable (morphological status, following segment, preceding segment, stress, frequency, phonological neighborhood density, voicing, and speech rate). The results for each variable will be discussed in terms of what human transcribers

said and in terms of how MFA classified the phones. FNs and TPs were instances in which a human transcriber listed the /t d/ as present, while FPs and TNs were cases in which a human transcriber listed the /t d/ as absent. FPs and TPs were instances in which MFA listed the /t d/ as present, while FNs and TNs were instances in which MFA listed the /t d/ as absent.

4.3.1 Morphological Status

In terms of morphological status, as shown in Figure 4.16, regular and irregular past tense words had more FNs and TPs (i.e., instances in which a human transcriber listed the word-final /t/ or /d/ as present as shown in Figure 4.16). Monomorphemes and other words had higher rates of FPs and TNs (i.e., where the transcriber listed a /t/ or /d/ as absent). This is in line with previous research that suggests regular past tense features less deletion (and monomorphemes feature more deletion). There is not a clear pattern with regards to MFA accuracy, as each morpheme status has approximately the same proportion of FPs and FNs compared to TPs and TNs. This is a logical result, as MFA is presumably insensitive to morphological properties.

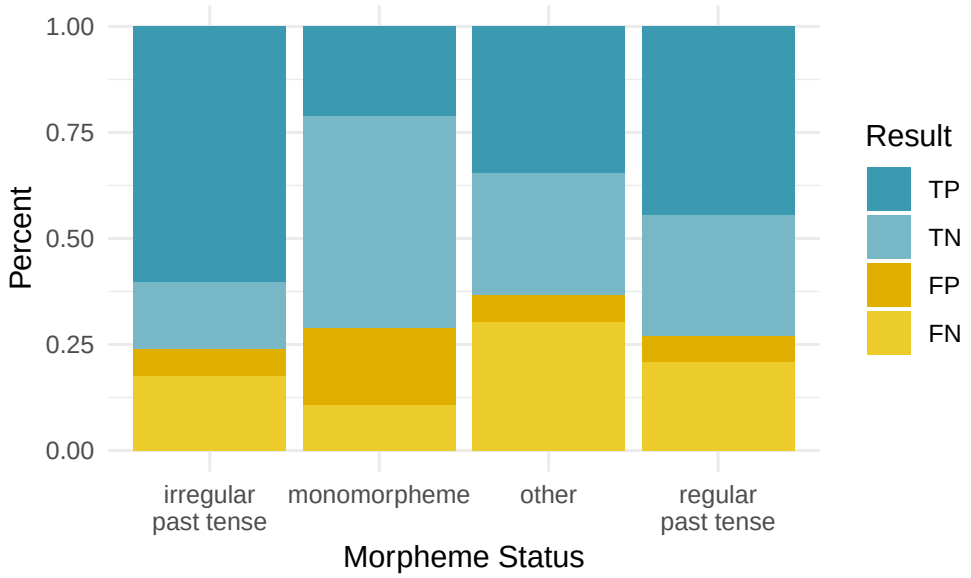

Figure 4.16: Effect of morpheme status

However, we can also compare F1 (the harmonic mean of precision and recall) in order to better understand the overall accuracy between different morphological classes. Figure 4.17 shows higher accuracy in irregular and regular past tense morphemes.

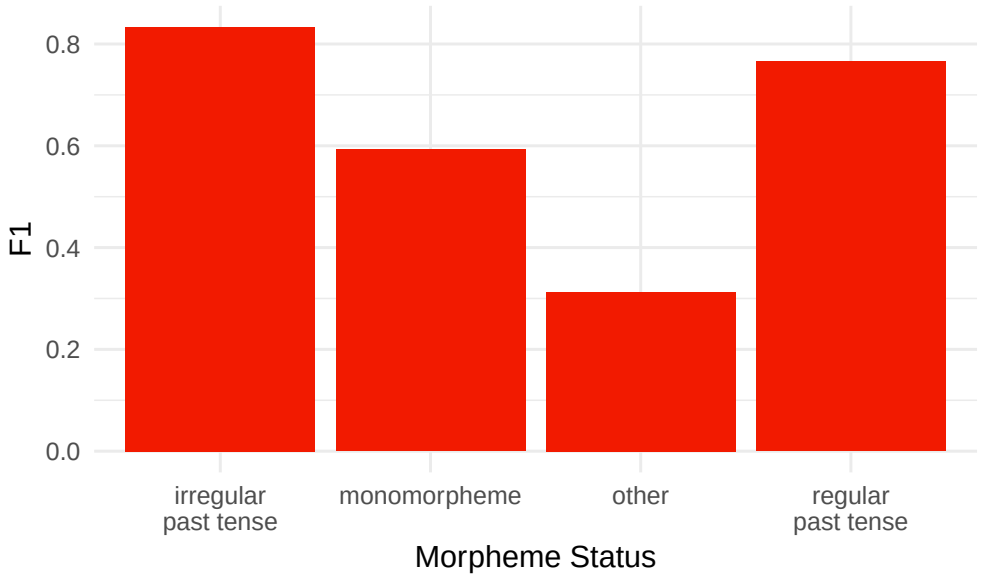

Figure 4.17: Effect of morpheme status, measured by F1

4.3.2 Following Segment

In terms of following segment, as shown in Figure 4.18, humans transcribed /t d/ deletion occurring most frequently when the following phone was a nasal, though affricates, stops, and fricatives do not exhibit higher rates of deletion, contradicting Yuan et al. (2020), who found the nasal /n/ to feature the 7th highest rate of deletion, with /tʃ, t, s, dʒ, d, ð/ featuring higher rates of deletion. The other segments have roughly the same ratio of FN and TP to TN and FP. Looking at TPs and TNs compared to FPs and FNs, MFA has higher accuracy when the segment proceeds a vowel or silence, as a vowel or silence is very different, acoustically, from a stop. On the other hand, MFA's accuracy is lower when the word-final /t d/ proceeds a stop or affricate. This appeals to intuition, as the acoustic model may map part of the acoustics of the following stop (or the stop portion of the affricate) to the word-final /t d/.

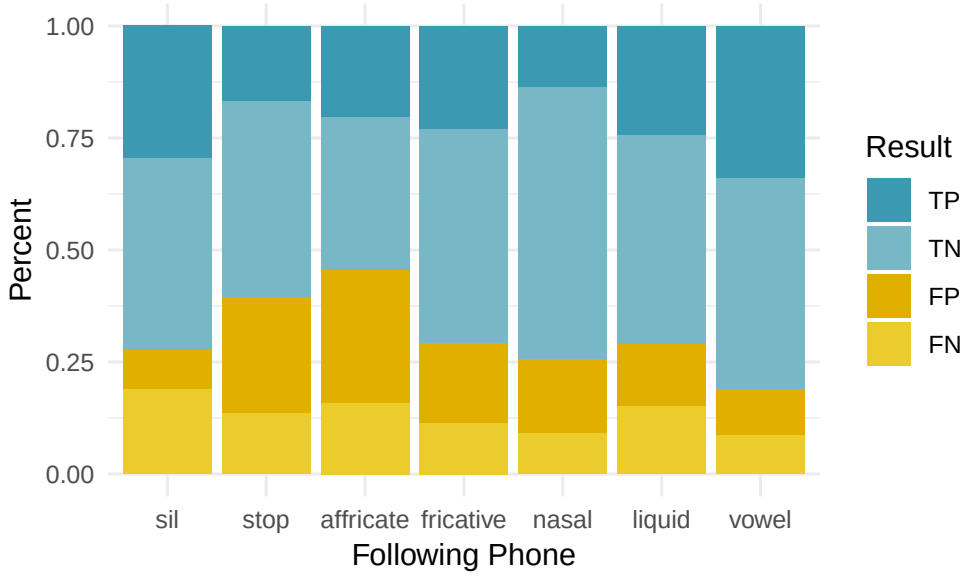

Figure 4.18: Effect of following phone

Again, we can see clearly that MFA succeeds more often when the stop occurs before a silence or vowel in Figure 4.19.

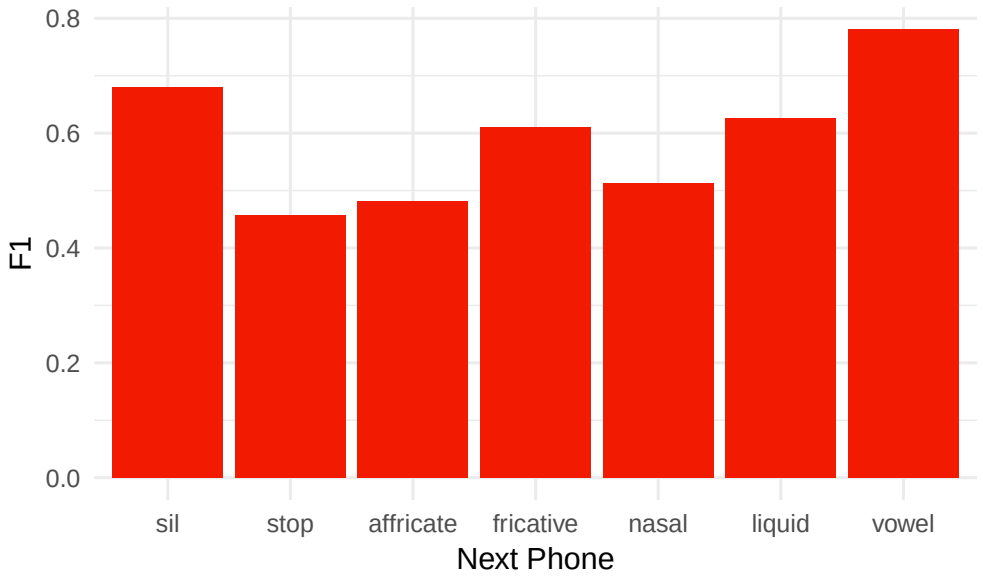

Figure 4.19: Effect of following phone, measured by F1

4.3.3 Preceding Segment

In terms of preceding segment, as shown in Figure 4.20, nasals as a preceding segment featured the highest levels of deletion (i.e., the highest rates of TNs and FPs) followed by fricatives, somewhat in line with Yuan et al. (2020), who found that the rates of deletion were greatest for /ng, jh, m, s, ch, n, z, sh/. Stops featured the lowest levels of deletion. MFA's accuracy is greatest for stops, followed by nasals, followed by fricatives, and is least accurate in affricates and liquids.

Additionally, as shown in Figure 4.21, stops, liquids, and affricates have the highest F1 score[23].

[23]In comparing the conclusions of Figure 4.21 to Figure 4.20, one may notice that nasals have the highest levels of accuracy in Figure 4.20 but a low F1 score in Figure 4.21. The formula for calculating F1 is as follows: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$, in other words, $F_1 = 2 \cdot \frac{(\frac{TP}{TP+FP}) \cdot (\frac{TP}{TP+FN})}{(\frac{TP}{TP+FP}) + (\frac{TP}{TP+FN})}$. The measurement of F1, as compared to accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$), is better when the cost of false negatives is too high (e.g., in medicine).

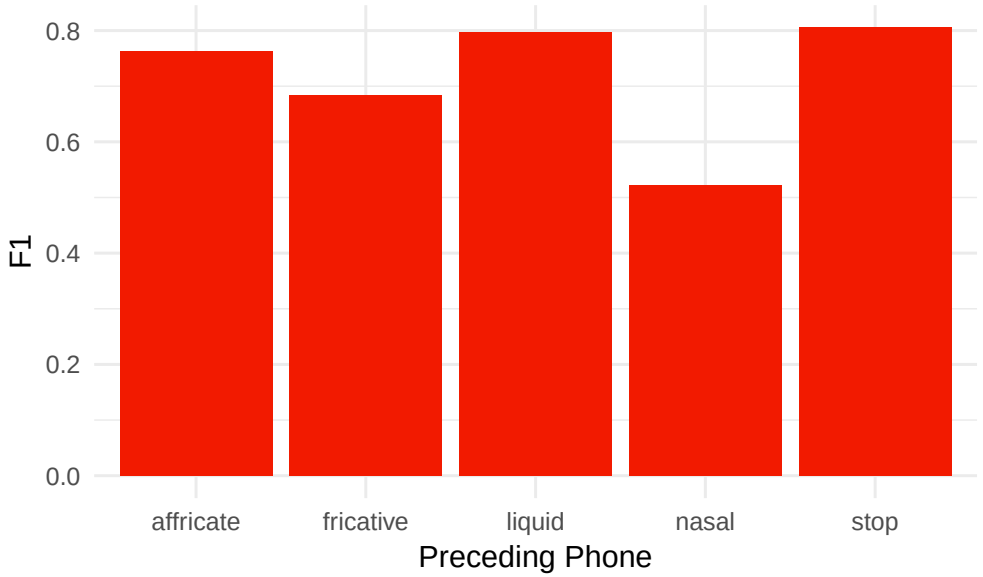

Figure 4.21: Effect of following phone, measured by F1

The effect of nasals could potentially be influenced by the frequency of *and*. Removing frequent, closed class words (*and, just, don't, kind* (as a part of *kinda*) *want, can't, around, doesn't, wasn't, sort,* (as a part of *sorta*) *haven't, won't, aren't, isn't, weren't, hasn't*) still shows higher deletion rates in human transcription (i.e., has more TNs and FPs) among /t d/ that follows a nasal in the remaining 9081 tokens, as shown in Figure 4.22.

The F1 score when the preceding phone is a nasal is much higher without these frequent, closed class words, as shown in Figure 4.23.

4.3.4 Stress

In terms of stress, as shown in Figure 4.24, /t d/ was deleted most often in unstressed syllables (in line with a great deal of research that unstressed syllables feature reduction, which can be thought of as a continuum with deletion at one end[24]). Syllables with primary stress and secondary stress have approximately the same rates of deletion. Interestingly, MFA

[24]See Chapter 1 for further discussion.

had greater rates of accuracy in unstressed syllables, shown by the greater number of TPs and TNs in Figure 4.24.

Looking at F1 in Figure 4.25, both primary and secondary stress had a higher F1 than unstressed.

4.3.5 Frequency

In terms of frequency, as shown in Table 4.5, FP and TP results (where humans transcribe /t d/ as deleted) have a higher frequency than TP and TN, meaning that more frequent words are more likely to feature deletion. Again, this is in line with a well-established line of research (e.g., Jurafsky et al. (1998), Jurafsky et al. (2001), Turnbull (2018)[25]) that concludes that higher frequency words are more reduced (in this case, deleted). Frequency does not play a role in MFA's accuracy. Like the result from morphological status, this appeals to intuition because MFA would not be sensitive to the frequency properties of words in the same way that a human listener would be.

Table 4.5: Effect of frequency

Result	**Mean measure of frequency**	**SD**
TP	5.48	0.99
TN	6.42	1.03
FP	6.67	0.97
FN	5.46	0.84

4.3.6 Phonological neighborhood density

In terms of phonological neighborhood density[26], as shown in Table 4.6, FPs and TNs have higher phonological neighborhood density, meaning that as phonological neighborhood density

[25]Though the relationship between frequency, predictability, and informativity, and the role that the later two may play, is not necessarily easy to tease apart. See Cohen Priva (2015) for more on informativity.

[26]106 values were removed because they had no phonological neighbors, as $log_{10}0 = -\infty$.

goes up, so does the likelihood of deletion, in line with the findings of Yuan et al. (2020). TNs and FNs, as compared to FPs and TPs, do not differ, suggesting that MFA is not sensitive to phonological neighborhood density.

Table 4.6: Effect of phonological neighborhood density

Result	**Mean measure of PND**	**SD**
TP	1.47	0.55
TN	1.68	0.44
FP	1.75	0.39
FN	1.41	0.54

4.3.7 Voicing

In terms of voicing, as shown in Figure 4.26, /d/ has very slightly higher rates of deletion (more TNs and FPs) than /t/, in line with Yuan et al.'s (2020) study. MFA's accuracy is also slightly higher for /d/.

Though the accuracy of /d/ was higher, Figure 4.26 shows the F1 of /t/ is higher.

4.3.8 Speech Rate

In terms of speech rate, as shown In Table 4.7, FPs and TNs compared with TPs and FNs do not show a difference. However, MFA was more likely to give /t d/ as absent in higher speech rate (whether it was accurate or not).

Table 4.7: Effect of mean speaking rate

Result	**Mean speaking rate in phones per second**	**SD**
FP	13.13	1.25
TP	13.05	1.29
TN	13.17	1.22
FN	13.24	1.22

4.3.9 Subphonemic Detail Analysis Results

As outlined in Subphonemic Detail, 10% of tokens from four subgroups (MFA judged /d/ to be present, MFA judged /d/ to be absent, MFA judged /t/ to be present, and MFA judged /t/ to be absent) were examined manually, and several subphonemic variables were recorded. The discussion of these results will begin with /d/ in subsubsection 4.3.9.1 and are followed by the results for /t/ in subsubsection 4.3.9.2.

4.3.9.1 /d/ Results

The results, in terms of whether /d/ was completely absent (no evidence of a closure or burst from /d/) or present in some form (a closure, burst, or both) are shown in Table 4.8.

Table 4.8: MFA vs. the author: /d/

		Human	
		∅	[t, d]
MFA	∅	710 (TN)	16 (FN)
	[t, d]	306 (FP)	171 (TP)

To judge inter-rater reliability between myself and MFA, we can use Cohen's kappa, which in this case is $\kappa = 0.38$, considered "fair" agreement (Landis & Koch, 1977). Now, we can examine each subphonemic component of the stop, starting with the closure (and the properties of the closure) in paragraph 4.3.9.1.1 followed by the burst and delayed release in paragraph 4.3.9.1.2 and paragraph 4.3.9.1.3 respectively. An additional component, the presence of excrescent stops (or in most cases, just bursts) will be discussed in paragraph 4.3.9.1.4. Finally, this section concludes with paragraph 4.3.9.1.5 that contains the results of a logistic regression that predicts MFA's behavior based on these subphonemic properties.

4.3.9.1.1 Closure Each token was examined to determine if a closure was absent or present. Overall, 1104 were judged to not have a closure and 99 were judged to have a closure. Figure 4.28 shows the counts of MFA's judgment within each category.

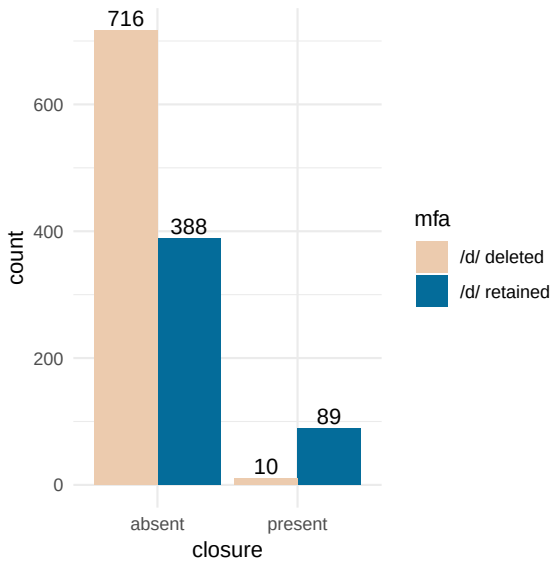

Figure 4.28: Closure of /d/ and MFA judgment

The proportion of MFA-judged stop presence is much higher in the tokens with the closure present, suggesting that closure may play a role in MFA's determination of /d/ absence or presence; the presence of a closure increases the likelihood that MFA will judge /d/ as present and the absence of a closure increases the likelihood that MFA will judge the /d/ as absent. Within tokens where the closure was judged to be present, we can look at the voicing of the closure (voiced vs. voiceless) and whether frication noise was present or absent during the closure. Overall, 60 tokens were judged to be voiceless, 32 were judged to be voiced, and seven were judged to be partially voiced. Figure 4.29 shows the proportion of MFA-judged stop presence in voiced and voiceless closures.

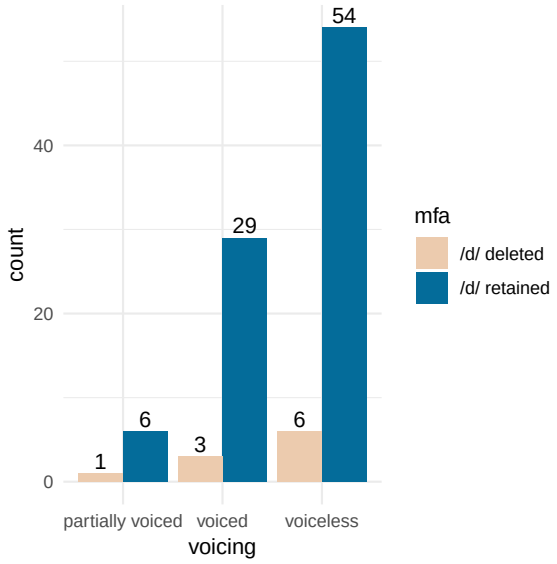

Figure 4.29: Closure voicing of /d/ and MFA judgment

Between all three designations (voiced, partially voiced, and voiceless), the proportion of MFA judging the stop as absent or present was approximately the same. Further, 13 tokens were judged to have frication in the closure and 86 were judged to not have frication in the closure. We can also visualize frication's role in MFA's judgment, shown in Figure 4.30.

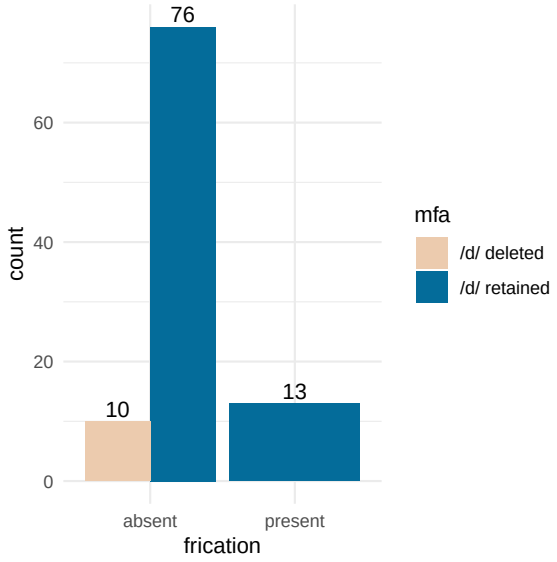

Figure 4.30: Closure frication of /d/ and MFA judgment

In tokens where frication was present, MFA judged the stop to be present 100% of the time. In tokens where the closure did not have frication, MFA judged about 25% of these to be absent. There is a difference, but it is unclear if frication influenced MFA due to the small number of tokens.

4.3.9.1.2 Burst Returning to the complete dataset of 1203 tokens, 163 were judged to have a burst, and 1040 were judged to not have a burst. In Figure 4.31, it is clear that the presence or absence of a burst corresponds to the presence or absence of the stop respectively, according to MFA.

Within the tokens that contained a burst, 134 had just a single burst and 32 had multiple bursts. Figure 4.32 shows that there is not a great deal of difference in MFA judgment between one and multiple bursts.

Within the tokens that contained a burst, 121 had a strong burst and 39 had a weak burst. Figure 4.33, similarly to Figure 4.32, shows that there is not a great deal of difference in MFA judgment between strong and weak bursts.

4.3.9.1.3 Delayed Release Overall, 49 tokens were judged to have a delayed release, meaning there was turbulent airflow following the burst. Figure 4.34 suggests that tokens with a delayed release were more likely to be judged as present by MFA.

4.3.9.1.4 Excrescent Stops As discussed in subsection 4.2.5, word-final /d/ deletion could give rise to conditions where an excrescent stop might be present. In the event that these conditions were present, and there was evidence of a burst, this was marked as "absent" (as the underlying /d/ wasn't present) and marked instead as excrescent. Overall, there were 28 excrescent stops, all of them consisting of only bursts (as the conditions that would give rise to an excrescent stop would also not be likely to have the closure of the stop). Figure 4.35 doesn't suggest a large difference between the excrescent stops, in terms of MFA judgment.

4.3.9.1.5 Logistic Regression To further understand the effects of these subphonemic properties of the acoustic signal on MFA, a binomial mixed effects logistic regression model with BOBYQA optimizer (Powell, 2009) was created using R package 'lme4' (Bates et al., 2015), using the formula shown in Equation 4.1.

$$\begin{aligned}\text{MFA response} \sim \text{closure} + \text{voicing} + \\ \text{delayed release} + \text{burst/excrescent burst} + \\ (1|\text{word}) + (1|\text{speaker})\end{aligned} \tag{4.1}$$

Initially, the model was made using all the subphonemic variables examined. However, this model would not converge, and only the variables shown in Equation 4.1 were included[27]. Additionally, variables were collapsed to accurately represent the acoustic signal, as burst and excresent burst were combined.

The results of this model are given in Table 4.9, made with the R package 'xtable' (Dahl et al., 2019). These results show that the presence of a closure, a delayed release, or a burst (excrescent or not) all influence MFA's decision.

A logistic regression model's estimate is given in log odds, and so a more interpretable measure, the probability that MFA judges /d/ as present, is calculated by $pr = \frac{e^{(x)}}{(1+e^{(x)}}$, where x is log odds. The probability of each feature is given in Table 4.10.

[27]The variable "frication" was removed due to the fact that it features complete separation, which occurs when "some values of a predictor [...] can perfectly predict the outcome", causing unreliability in logistic regression (Levshina, 2015, p. 273). The presence of frication perfectly predicts MFA's judgment. Even when conflating the closure variable and frication variable, there is still complete separation. Therefore, the decision was made to exclude this variable in the logistic regression model. Additionally, burst number and burst strength displayed quasi-complete separation, and they were also excluded from the model.

Table 4.9: Mixed logistic regression model for /d/

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.75	0.16	-4.75	2.07e-06 (***)
closure = present	1.48	0.51	2.94	0.00 (**)
voicing = partially voiced	-2.04	1.31	-1.55	0.12
voicing = voiced	0.36	0.84	0.44	0.66
delayed release = present	1.64	0.83	1.98	0.05 (*)
burst = present	2.00	0.24	8.19	2.54e-16 (***)

Table 4.10: Probability of significant features in /d/ logistic regression

	Probability of MFA judging tokens as present
Closure	81.5%
Delayed release	83.8%
Burst	88.1%

4.3.9.2 /t/ Results

The results, in terms of whether /t/ was completely absent (no evidence of a closure, burst, or variant of /t/) or present in some form (a closure, burst, or variant) are shown in Table 4.11.

Table 4.11: MFA vs. the author: /t/

		Human	
		∅	[t, d]
MFA	∅	522 (TN)	140 (FN)
	[t, d]	156 (FP)	328 (TP)

To judge inter-rater reliability between myself and MFA, we can use Cohen's kappa (Cohen, 1960), which in this case is $\kappa = 0.47$, considered "moderate" agreement (Landis & Koch, 1977). The rest of this section will examine each subphonemic component of the stop, starting with the closure and its voicing and frication properties; the burst and its properties,

the number and strength; the presence of a delayed release, excrescent bursts, glottalization, and flapping.

4.3.9.2.1 Closure Overall, 946 tokens were judged to not have a closure, and 200 were judged to have a closure. As shown in Figure 4.36, MFA was more likely to judge a /t/ as present if a closure was present, and less likely to do so if a closure was absent.

Within tokens that had a closure, 43 were voiced, 148 were voiceless, and nine were partially voiced. Figure 4.37 shows approximately equal proportion of MFA's judgment of the token as present or absent within each voicing category, suggesting that MFA's judgment is not affected by this acoustic characteristic.

Within tokens that had a closure, 22 had frication, and 178 did not have frication. Similarly to Figure 4.37, Figure 4.38 suggests that frication in the closure does not affect MFA's judgment, as the proportions of MFA's judgment of present or absent remain roughly the same across the two categories.

4.3.9.2.2 Burst Of the 1146 tokens of /t/, 146 tokens were judged to have a burst. Of these 146, 118 had one burst, and 28 had multiple bursts. Further, 98 featured strong bursts (energy present throughout the spectrum), while 48 were weak (energy not present throughout the spectrum). There is a difference in MFA judgment between the two burst categories (present and absent) shown in Figure 4.39, as MFA was more likely to judge the /t/ as deleted if the burst was absent, and more likely to judge the /t/ as present if the burst occurred.

However, the properties of the burst (number, shown in Figure 4.40 and strength, shown in Figure 4.41) do not show large differences in MFA's judgment in the categories.

4.3.9.2.3 Delayed Release Sixty-eight tokens featured delayed release, and separately, 75 tokens feature delayed release after /s/[28]. As shown in Figure 4.42, there is a large

[28]These were classified separately due to their special status in Olive et al. (1993). These characteristics were discussed in subsection 4.2.5.

difference in MFA judgment between tokens featuring delayed release and those that do not, with tokens featuring a delayed release almost always classified as present by MFA.

A similar, yet less pronounced, pattern is shown in delayed release tokens after /s/, in Figure 4.43.

4.3.9.2.4 Excrescent bursts Overall, there were 16 excrescent bursts. In these 16, there is a difference in MFA judgment between tokens featuring an excrescent stop and those that did not, as shown in Figure 4.44. If MFA judged the /t/ as absent, it was more likely to not feature an excrescent burst.

4.3.9.2.5 Flaps There were 19 realizations of /t/ as a flap, [ɾ], for example, in *sort of*. From the visualization shown in Figure 4.45, there is not a great difference in MFA judgment between the two groups (flapped vs. not flapped).

4.3.9.2.6 Glottalization Within the /t/ data, 130 tokens were judged to feature glottalization. Again, according to the visualization in Figure 4.46, there is not a great deal of difference in MFA's judgment between the two groups.

4.3.9.2.7 Logistic Regression As done in the analysis of /d/, a logistic regression was created for /t/ data, with MFA's judgment as the dependent variable, and the acoustic characteristics above as independent variables. Both acoustic characteristics of the burst (number and strength) caused the model to be rank deficient, and these predictors were taken out. The voicing variable, with three levels (voiced, voiceless, and partially voiced) featured complete separation. Therefore, "partially voiced" was recoded as "voiced".

The model included the formula given in Equation 4.2, and the model results are shown in Table 4.12

$$\begin{aligned}\text{MFA response} \sim\ & \text{closure} + \text{voicing} + \text{closure frication} + \\ & \text{burst/excrescent burst} + \text{delayed release} + \\ & \text{delayed release after sibilant} + \\ & \text{flapping} + \text{glottalization} + \\ & (1|\text{word}) + (1|\text{speaker})\end{aligned} \tag{4.2}$$

Table 4.12: Mixed logistic regression model for /t/

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.85	1.11	-1.66	0.10 (.)
closure = present	3.34	0.40	8.26	< 2e-16 (***)
voicing = voiced	0.39	0.61	0.64	0.52
frication = present	0.24	0.85	0.29	0.78
burst = present	0.74	0.35	2.14	0.03 (*)
delayed release = present	1.55	0.82	1.88	0.06 (.)
delayed release after /s/ = present	1.61	0.28	5.77	7.82e-09 (***)
flap = present	1.09	0.59	1.85	0.06 (.)
glottal = present	1.22	0.22	5.45	5.06e-08 (***)

Like /d/, the presence of a closure was significantly more likely to result in MFA to judge the /t/ as present; 97% more likely when the closure was voiced (or partially voiced). Similarly to /d/, the presence of a burst was 68% more likely to result in MFA judging the /t/ as present, and a delayed release was 83% more likely to result in MFA judging /t/ as present. The other three variations of /t/ recorded were all significant. The noise/delayed release after /s/ resulted in an 83% increase, the flap a 75% increase, and the glottalization a 77% increase in the likelihood that MFA judged /t/ as present. These probabilities are summarized in Table 4.13.

Table 4.13: Probability of significant features in /t/ logistic regression

	Probability of MFA judging tokens as present
Closure	96.5%
Burst	67.7%
Delayed release	82.5%
Delayed release after /s/	83.3%
Flap	74.8%
Glottal	77.2%

4.4 Discussion

This discussion section is divided into two parts: subsection 4.4.1 consists of a brief summary of where deletion is likely to be present in the Buckeye Corpus, according to human transcribers, and subsection 4.4.2 is a discussion of MFA performance after dictionary modification.

4.4.1 /t d/ Deletion in the Buckeye Corpus

Investigating the variables by separating into TP/FP/FN/TN allowed the simultaneous investigation of /t d/ deletion rates (as human-labeled) in addition to investigating MFA performance of marking /t d/ as present or absent. Table 4.14 shows a summary of where deletion is more likely to occur according to the human annotators of the Buckeye Corpus.

4.4.2 MFA Performance

This subsection discusses MFA performance in two ways: subsubsection 4.4.2.1 discusses how the variables measured affect MFA's accuracy, and subsubsection 4.4.2.2 discusses how the agreement between MFA and a human transcriber agree with that of two human transcribers.

Table 4.14: /t d/ deletion in the Buckeye Corpus according to human annotators

Variable	**Rate of Deletion**
Morpheme status	monomorpheme/other > regular/irregular past tense
Following segment	nasal/liquid > stop/affricate/fricative > sil/vowel
Preceding segment	nasal > affricate/fricative/liquid > stop
Stress	unstressed > primary stress > secondary stress
Frequency	more frequent > less frequent
PND	more neighbors > less neighbors
Voicing	d > t
Speech rate	effect unclear

4.4.2.1 Effect of Variables on MFA Performance

In forced alignment, as described previously, an acoustic model is used to locate optimal alignment. Therefore, it follows that the variables that might affect MFA's performance are ones dealing with *acoustics* rather than any morphological variable present. To see which variables affected MFA's accuracy, we can look at rates of FPs and FNs, in addition to the statistical model that was built. Differences in rates of FPs and FNs are shown in Table 4.15. Only the following phone and stress of the syllable influence MFA's accuracy. This appeals to the logic that non-acoustic variables would not influence MFA performance. Within the variable preceding phone, rates of false positives were lower, and rates for false negatives were lower, when the preceding sound was a stop. It appeals to intuition that the acoustics of the preceding stop could potentially be misaligned as the acoustics for the word-final /t d/. Within stress, MFA was more likely to judge a token as a FP for primary stress, and judge a token as FN for secondary stress. MFA's greatest accuracy rates were with unstressed tokens.

4.4.2.2 Accuracy of MFA Compared to Human Transcribers

As mentioned in the Results section, the agreement in the current study between MFA and the human transcription of the Buckeye Corpus was 71%. In previous studies on inter-

Table 4.15: MFA FNs by variable

Variable	**Rate of FNs and FPs**
Morpheme status	no difference
Following phone	affricates/stops > fricatives/liquids > sil/nasal/vowel
Stress	stressed > unstressed
Frequency	no difference
PND	no difference
Voicing	no difference

transcriber agreement on the Buckeye Corpus, phone matching between transcribers was only 76% (Raymond et al., 2002) or 80% (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005). In Raymond et al.'s (2002) study, transcribers agreed "on the existence of a phone (though not necessarily on its label, that is) for 86% of the events identified as a phone by at least one transcriber" (2002, p. 2), but stops were agreed on only approximately 70% of the time, as shown in Figure 4.47.

A more recent study also showed that the percent of stops agreed on unanimously by six transcribers is 74% (Pitt et al., 2005). Raymond et al. (2002) also state that of 43 disagreements that resulted in a difference of word transcribed, the most of any group was final segment deletion, giving the example of [fam] vs. [famd] (Raymond et al., 2002). The present study's results in combination with Raymond et al.'s (2002) and Pitt et al.'s (2005) result suggest that MFA, with dictionary modifications, is on par with inter-transcriber agreement. If an out-of-the-box dictionary was used with MFA, where deletion is not an option, it follows that TNs would be reclassified as FPs, and FNs would be reclassified as TPs. The result would then be 16,796 FPs and 6726 TPs, a 28.5% accuracy rate. Overall,

these additions to the dictionary resulted in a 149.1% increase in accuracy[29] and a 71.3% decrease in false positives.

Looking at the results from the manual subphonemic analysis in subsection 4.3.9, MFA's judgment of /d/ is clearly linked to the presence of a burst, as hypothesized. Additionally, closure and delayed release were present, these being the three "components" of a stop. However, the results for /t/ were a bit more diffuse. While the closure, burst, and delayed release of a canonical stop increased the likelihood of MFA judging /t/ as present, the variants did as well. This would suggest that the acoustic model of MFA is sensitive to the variability in the realization of /t/. My hypothesis that the burst would be influential was right, though there were more significant predictors than I hypothesized.

Overall, with a fair Cohen's kappa (Cohen, 1960) value for inter-transcriber ($\kappa = 0.38$) reliability for /d/ combined with the lower number of FN as compared to FP in addition to the moderate Cohen's kappa (Cohen, 1960) value for /t/ ($\kappa = 0.47$) seems somewhat encouraging for this methodology as a tool for sociophonetic research. The results of this research also provide an interesting look into the "black box" of MFA, examining the acoustic correlates of what features MFA uses to provide phone-level alignment. In order to improve upon MFA's accuracy in alignment, this kind of research will need to be done.

4.4.3 Future Work

This section outlines three areas for future work in this topic: modeling other aspects of variation, modification of HMMs, and the use of weighted finite state transducers (WFSTs). To start, the initial results of this dictionary modification are somewhat promising, and additional variation, such as flapping, assimilation, etc., may be able to be captured by MFA with dictionary modifications by simply adding entries to model pronunciation variation.

[29]Using the formula accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+TN+FN}}$, accuracy without dictionary modifications is 28.5% ($\frac{6726}{19{,}522}$), with dictionary modifications is 85.8%($\frac{16{,}756}{19{,}522}$). Calculating percent change of accurate tokens using the formula $C = \frac{x_2 - x_1}{x_1} * 100$, where C refers to percent change, x_2 refers to the final value, and x_1 refers to the initial value, means that there is a 149.1% increase in accuracy.

For example, the canonical pronunciation of "butter": B AH T ER (as given in the CMU Pronouncing Dictionary (Weide, 1998)) could be given in the dictionary as its realized form, B AH DX ER, as well. Another study that examines modifying the dictionary pronunciation in ASR is Vasilescu et al. (2019), in which the deletion of *-l*, the Romanian definite article, is modeled in the ASR system.

An alternative to dictionary modification would be to modify the HMMs themselves. According to Yuan, Lai, Cieri, and Liberman (2018), alignment of segments that undergo deletion and temporal reduction can be improved by using skip-state HMMs. These HMMs have transitions that allow them to skip over the three states (beginning, middle, and end) of a triphone model to allow for phones that have been deleted. An example of an HMM with skip-state transitions is shown in Figure 4.48. In fact, Yuan et al. (2020) uses skip-state HMMs using HTK and achieved an accuracy rate of 79.1% as opposed to an accuracy rate of 73.5% with modification of the dictionary to identify /t d/ deletion. Despite this work that shows skip-state HMMs to produce more accurate alignment than dictionary modification[30], I argue that improvement from dictionary modification is useful to sociolinguistic inquiry due to the ease of implementation and the low rate of false positives.

Another area for future inquiry is how a weighted finite state transducer might improve the dictionary. As the lexicon is modeled with finite state transducers we could also use *weighted* finite state transducers to model the probability of one entry in the dictionary over another (see Mohri, Pereira, and Riley (2002)). For example, the current study and prior literature found regular past tense to be deleted roughly 25% of the time. An example of a weighted finite state transducer for the word *walked*, shown in Figure 4.49 would be able to specify the variant [wakt] is 75% likely to occur while [wak] is only 25% likely to occur. It is important to point out, however, that this would depend on knowing the "actual" rate of /t d/ deletion in the variety being modeled, which is difficult to do.

[30]Ultimately, neural nets had the highest accuracy, 86.7%.

4.4.4 Limitations

One limitation is that, by its nature, transcription is based on *segments*, not acoustics, a point also also acknowledged by Schuppler (2011). Even despite my best efforts to categorize the subphonemic properties of these tokens, it is still a categorical measure of reduction, which really is a continuous process. Furthermore, neural net forced aligners may be the most accurate, a point that is acknowledged earlier in the chapter, but ultimately these techniques may not widely accessible by sociolinguists. Additionally, one issue with adding pronunciation variants to the dictionary is the potential to add homophones that give rise to error (Adda-Decker & Lamel, 1999).

4.5 Conclusion

This chapter investigates the effect of dictionary modification in MFA and subphonemic properties of /t d/. The Librispeech dictionary was modified in order to include variant pronunciations for words that have word-final consonant clusters that end in /t d/, as they are more likely to undergo word-final /t d/ deletion. Overall, MFA's accuracy was 71%, performing at a level near human inter-transcriber agreement. This result is significant, because it shows that sociolinguistic variation can be incorporated into forced alignment, meaning more accurate alignment (and therefore, more accurate measurements) in large-scale sociophonetic studies. Analysis of subphonemic variation revealed that /d/ tokens containing a burst were likely to be classified as present, while /t/ tokens containing a closure, burst, or another variant of /t/ were likely to be judged as /t/. This is significant because it provides an acoustic explanation for MFA's results.

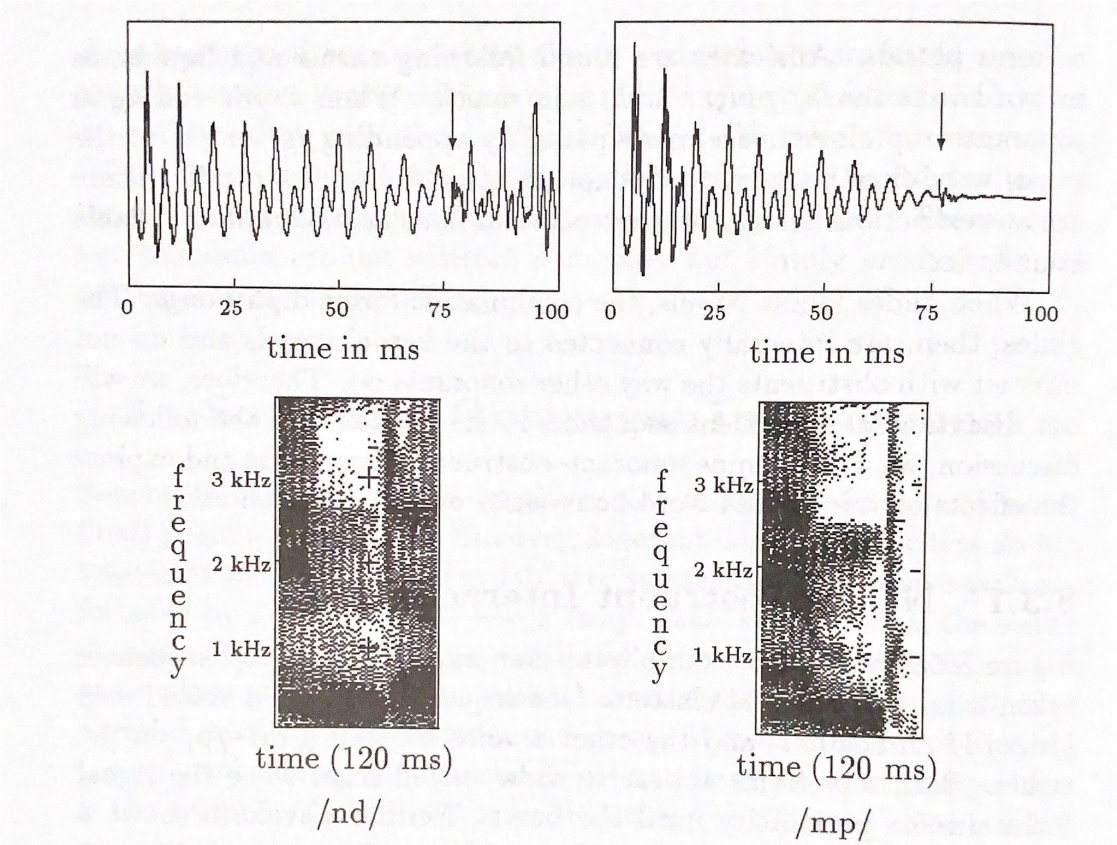

Figure 8.55: Nasal–stop clusters

Figure 4.13: /nd/ cluster, reproduced from Olive, Greenwood, and Coleman (1993, p. 296)

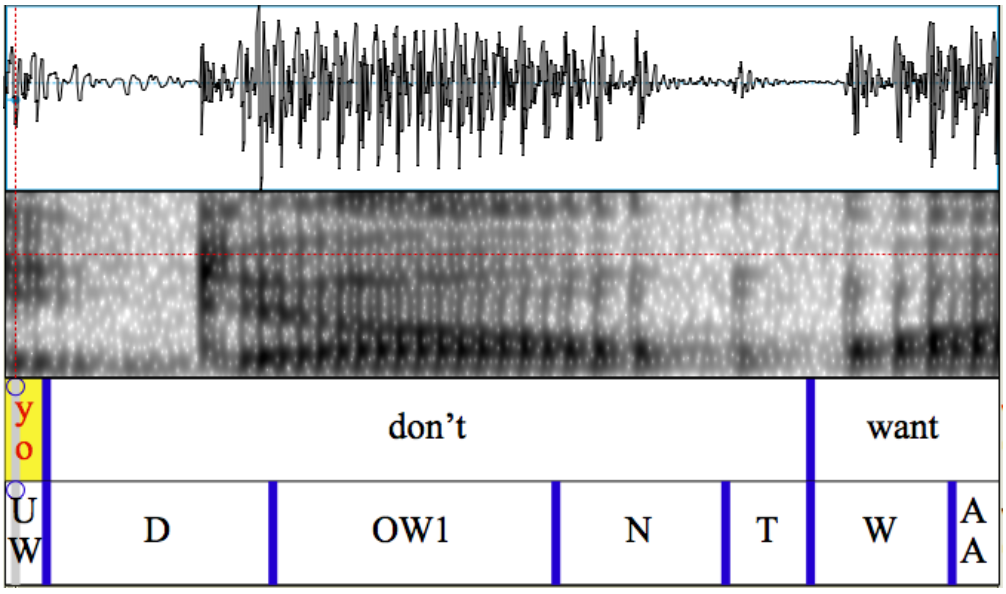

Figure 4.14: Glottalized /t/

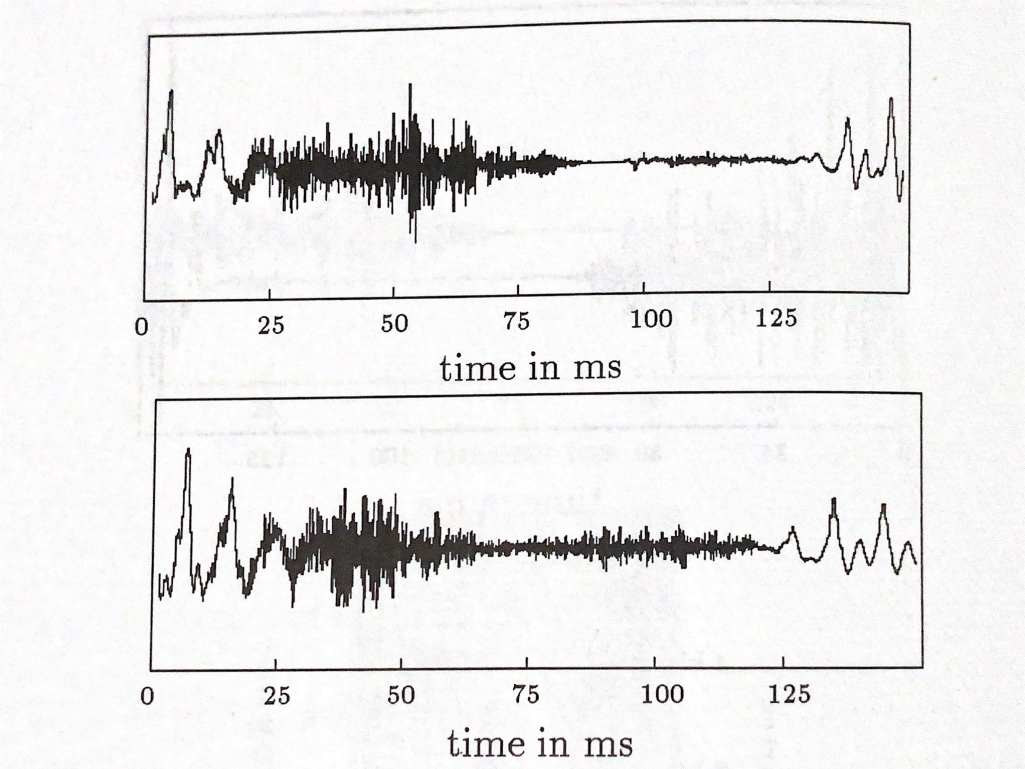

Figure 8.29: Two instances of word-final /st/ clusters

Figure 4.15: /-st/ cluster, reproduced from Olive, Greenwood, and Coleman (1993, p. 260)

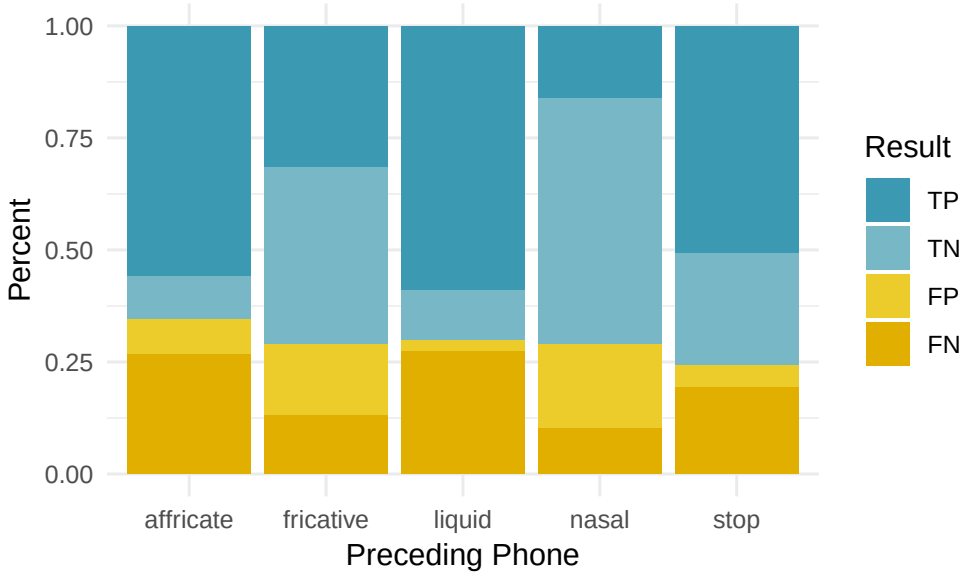

Figure 4.20: Effect of preceding phone

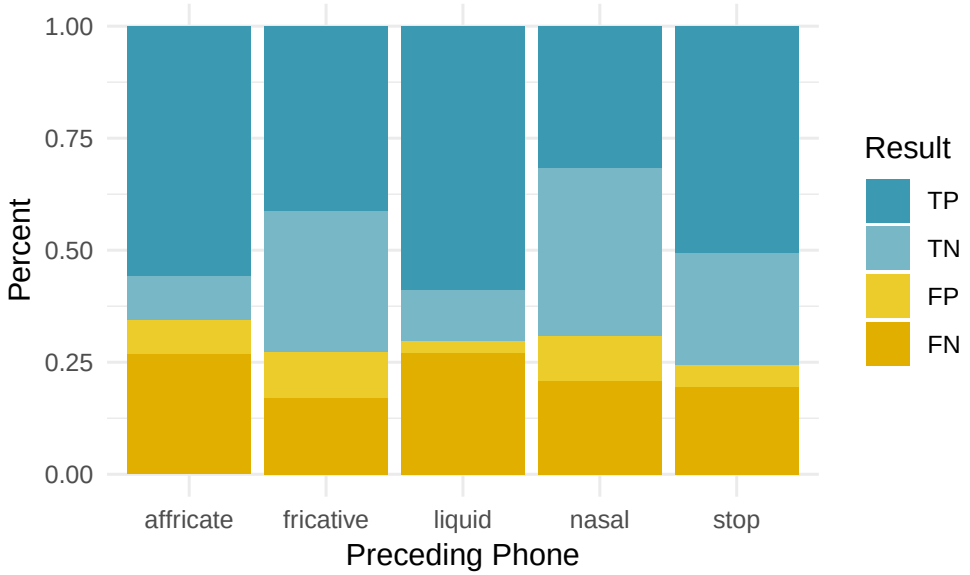

Figure 4.22: Effect of preceding phone without frequent words

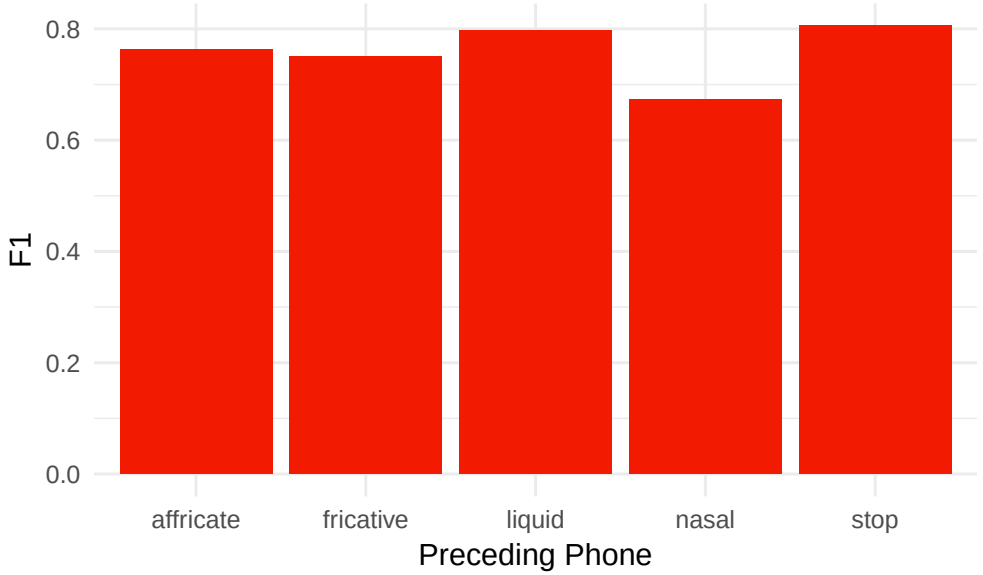

Figure 4.23: Effect of preceding phone without frequent words, measured by F1

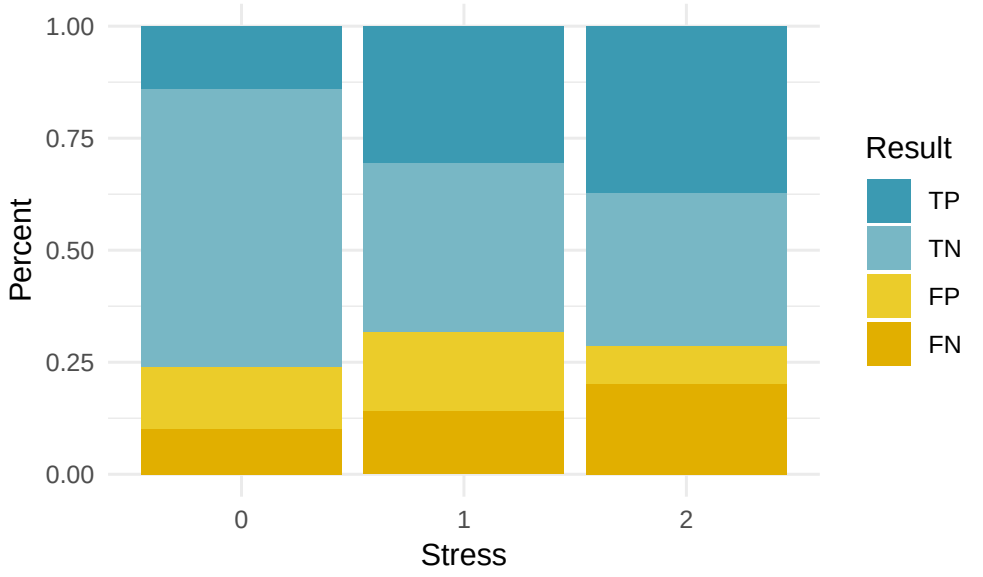

Figure 4.24: Effect of stress

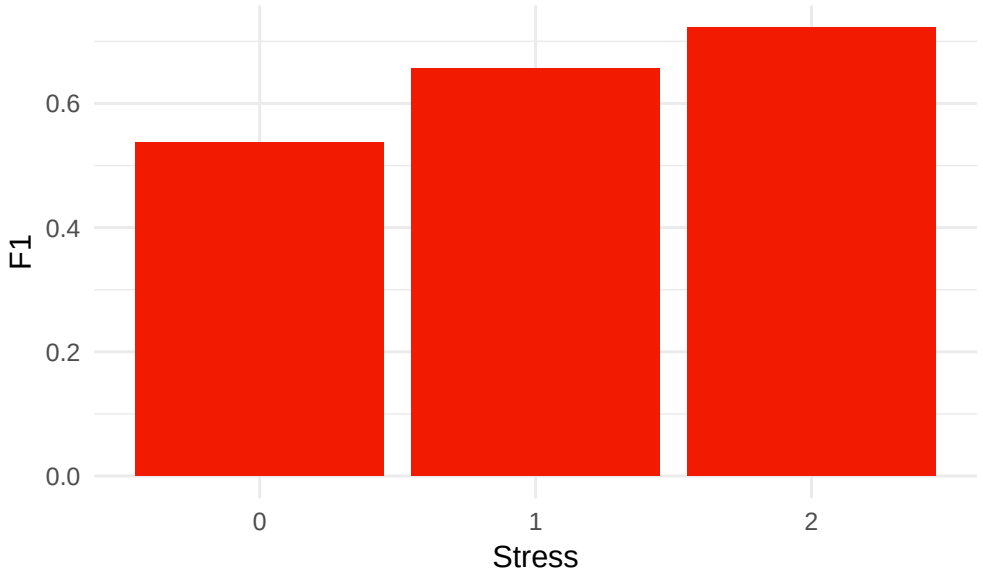

Figure 4.25: Effect of stress, measured by F1

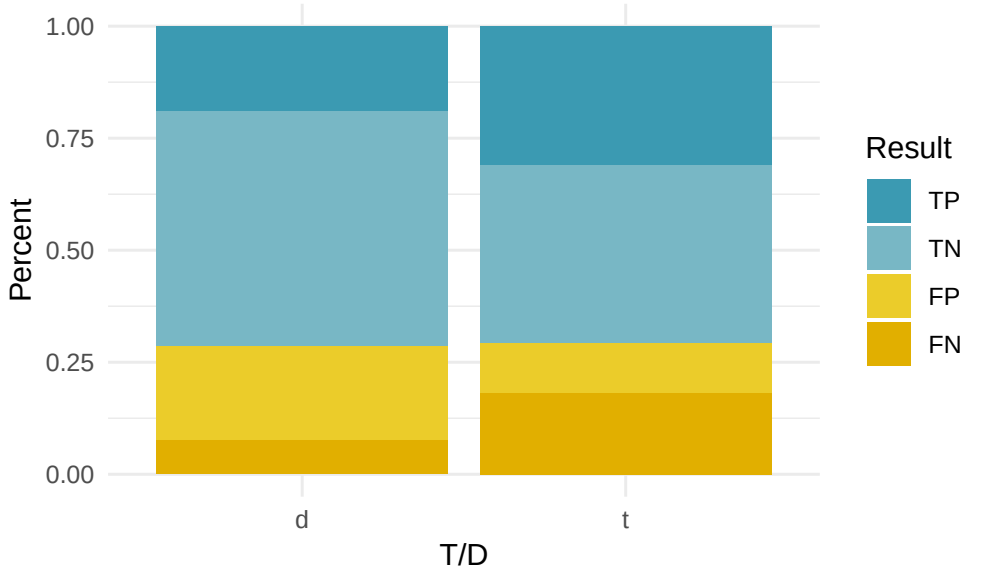

Figure 4.26: Effect of voicing

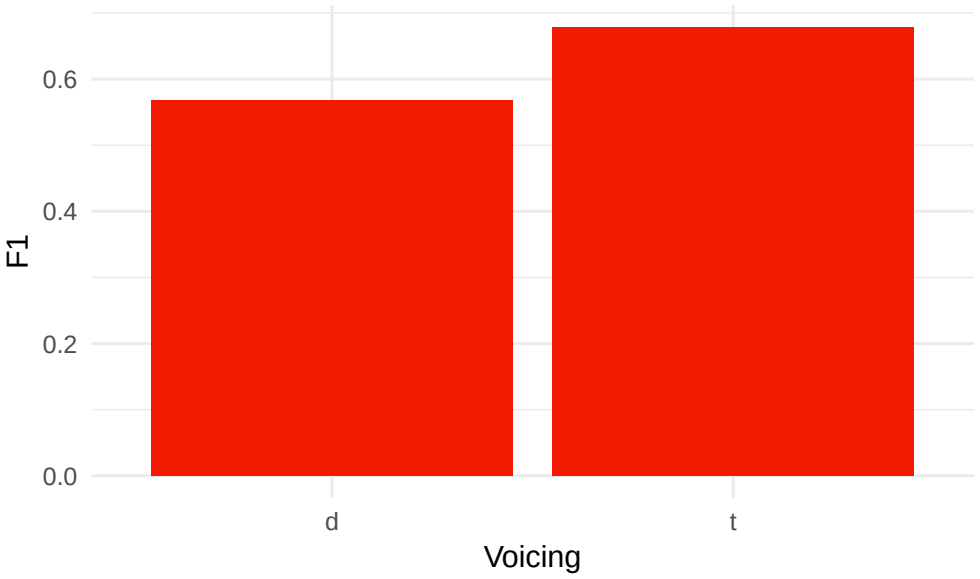

Figure 4.27: Effect of voicing, measured by F1

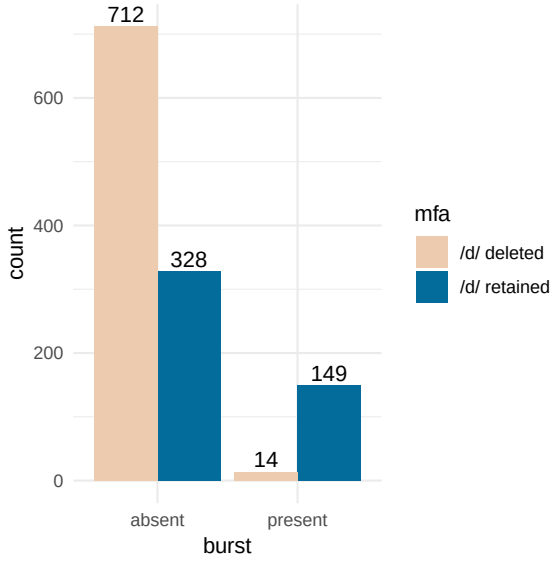

Figure 4.31: Burst of /d/ and MFA judgment

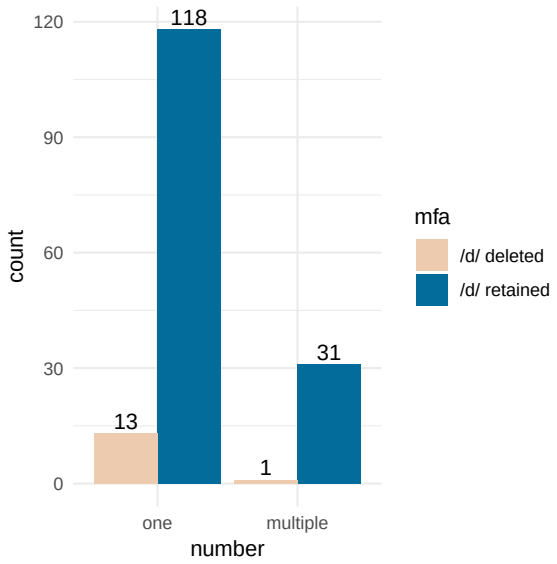

Figure 4.32: Burst number in /d/ and MFA judgment

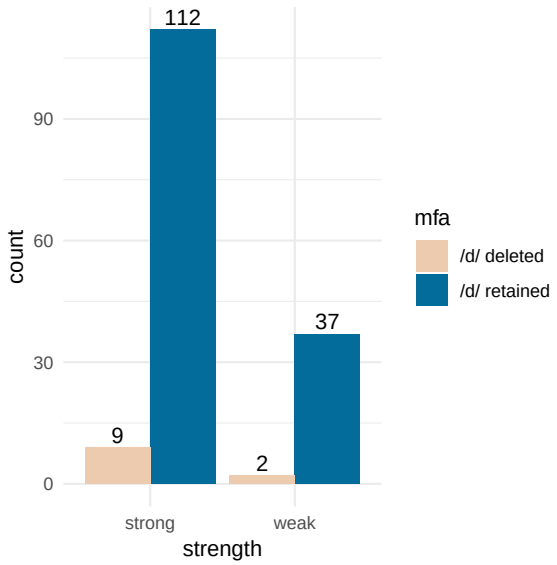

Figure 4.33: Burst strength of /d/ and MFA judgment

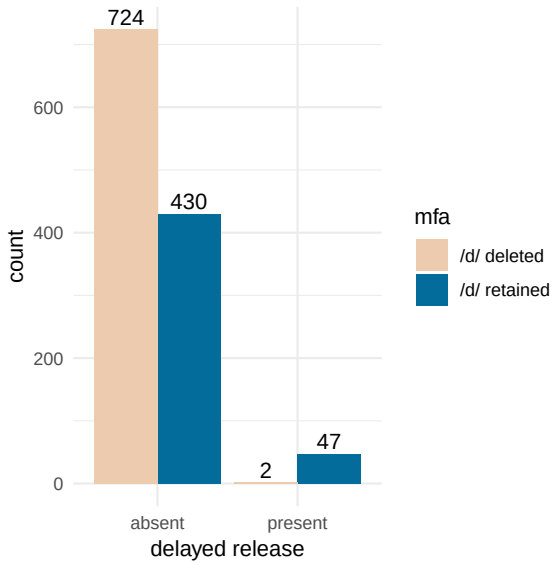

Figure 4.34: Delayed release of /d/ and MFA judgment

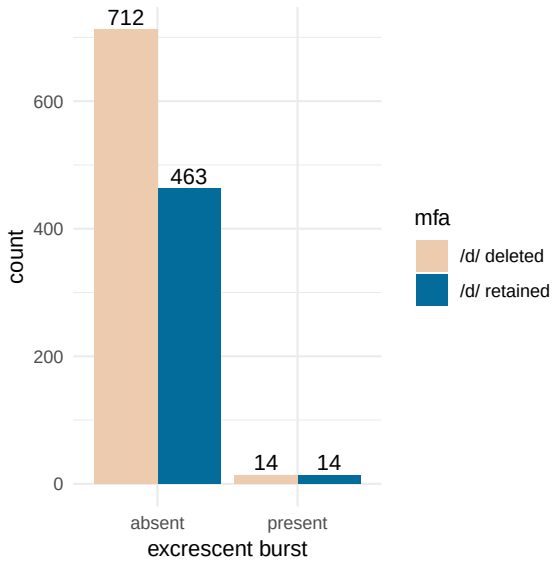

Figure 4.35: Excrescent bursts from /d/-deletion and MFA judgment

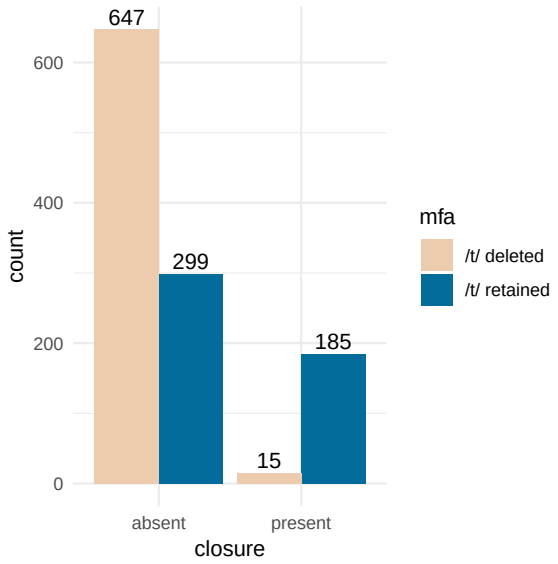

Figure 4.36: Closure of /t/ and MFA judgment

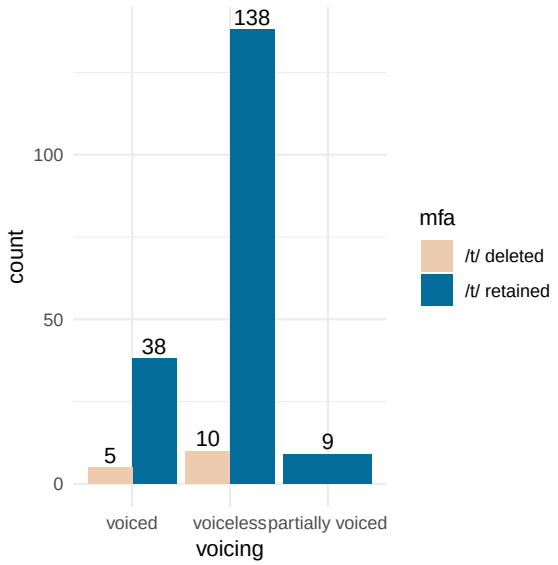

Figure 4.37: Closure voicing of /t/ and MFA judgment

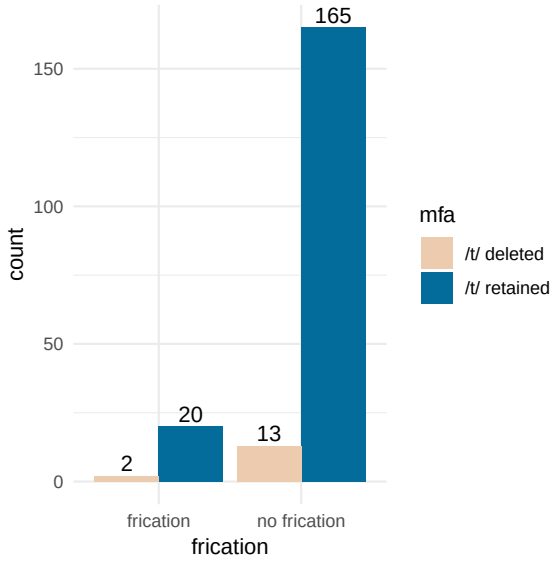

Figure 4.38: Closure frication of /t/ and MFA judgment

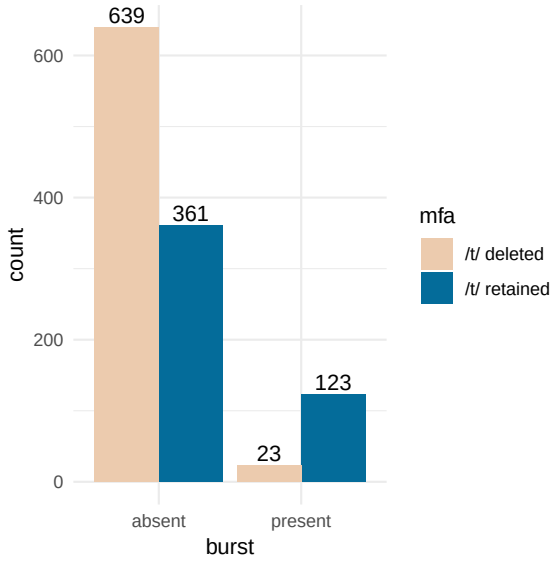

Figure 4.39: Burst of /t/ and MFA judgment

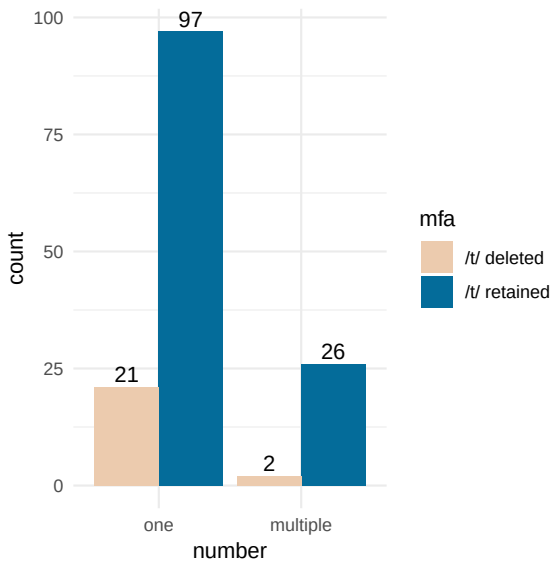

Figure 4.40: Burst number in /t/ and MFA judgment

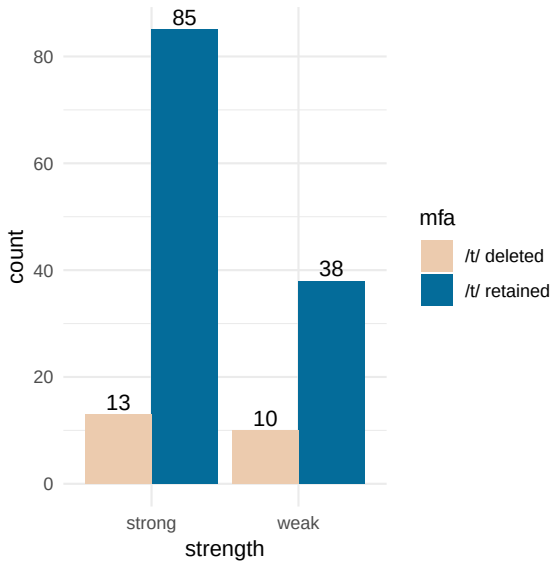

Figure 4.41: Burst strength of /t/ and MFA judgment

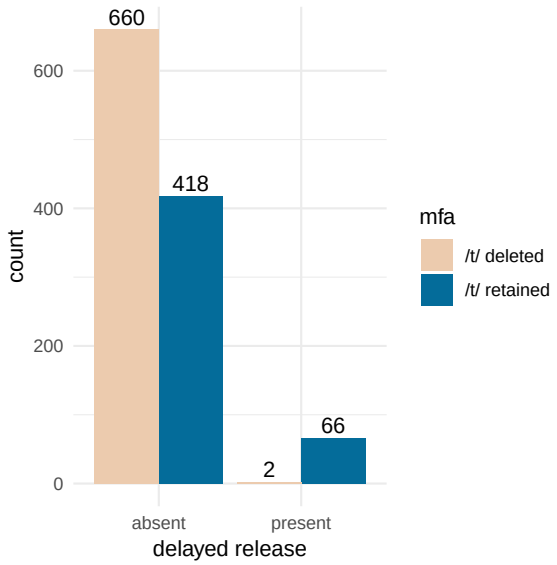

Figure 4.42: Delayed release of /t/ and MFA judgment

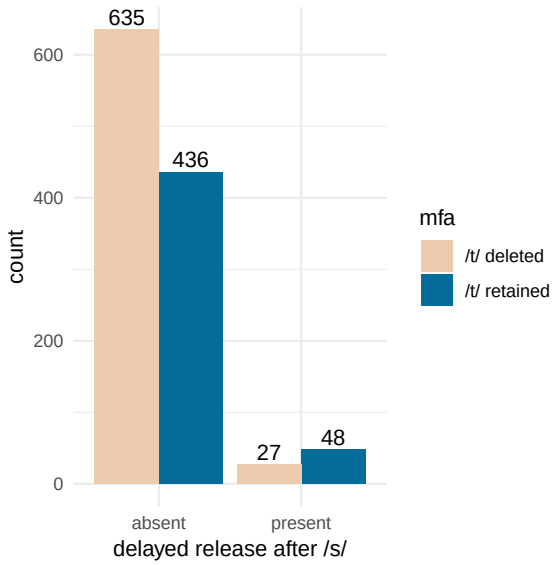

Figure 4.43: Delayed Release and MFA judgment

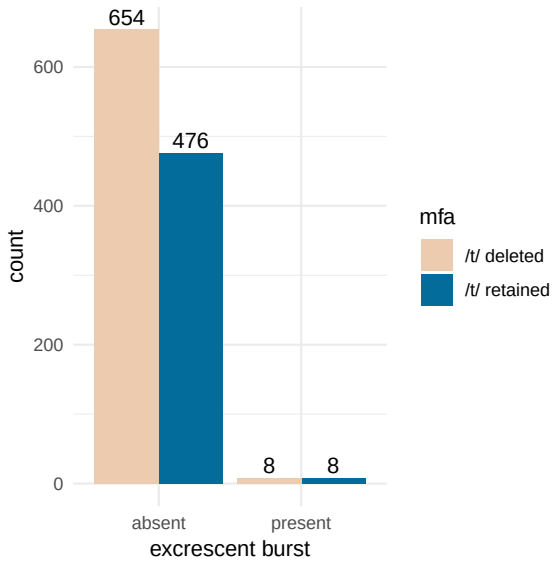

Figure 4.44: Excrescent bursts from /t/-deletion and MFA judgment

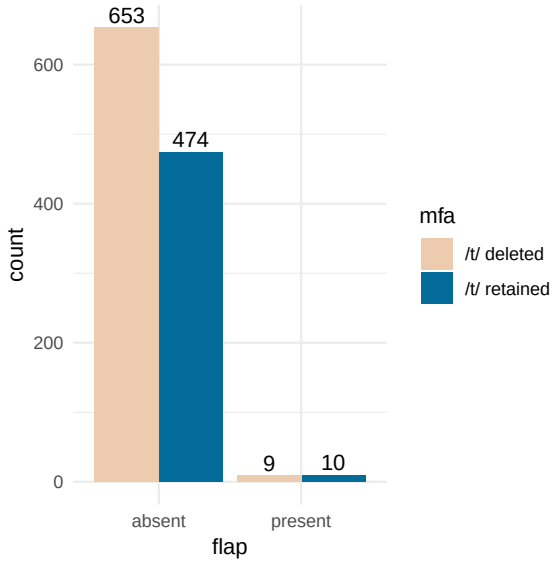

Figure 4.45: Flapped /t/ and MFA judgment

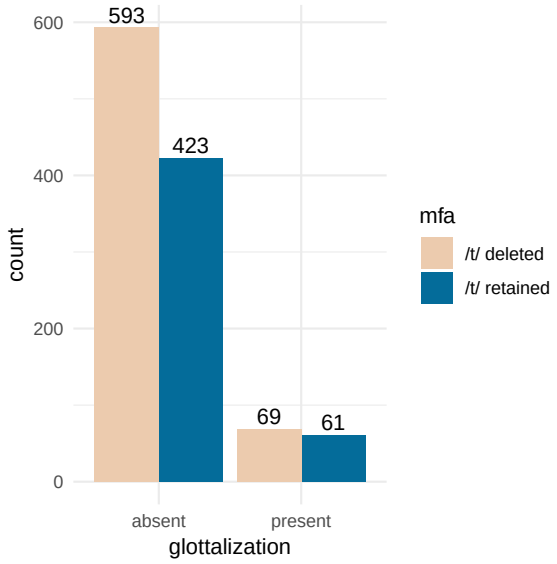

Figure 4.46: Glottalized /t/ and MFA judgment

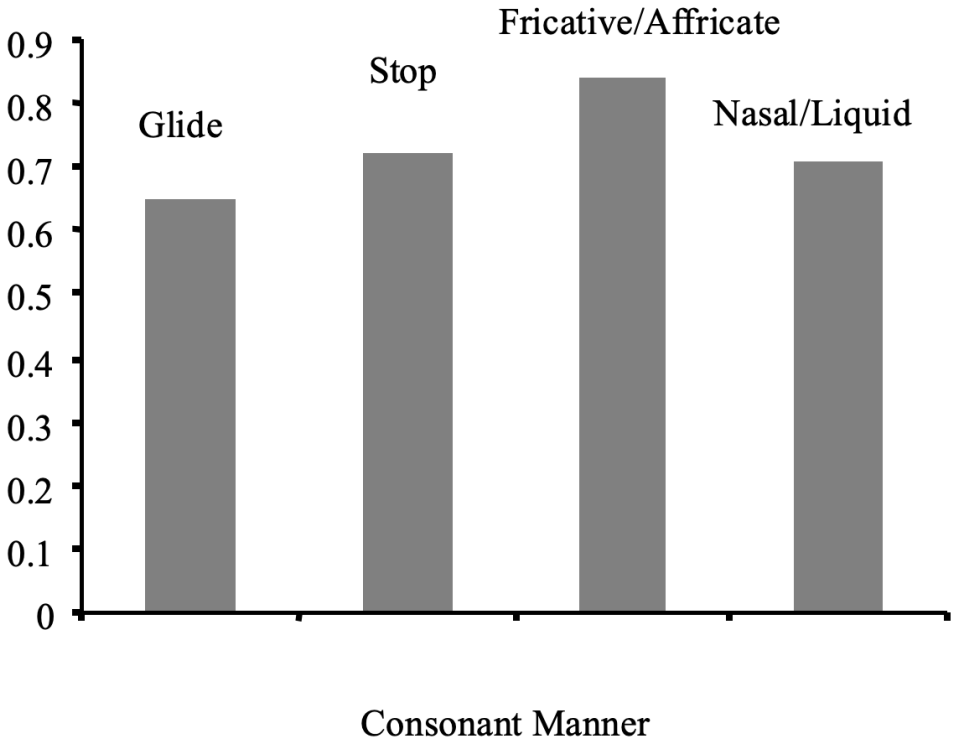

Figure 4.47: Consonant label agreement by manner, reproduced from Raymond et al. (2002, p. 2)

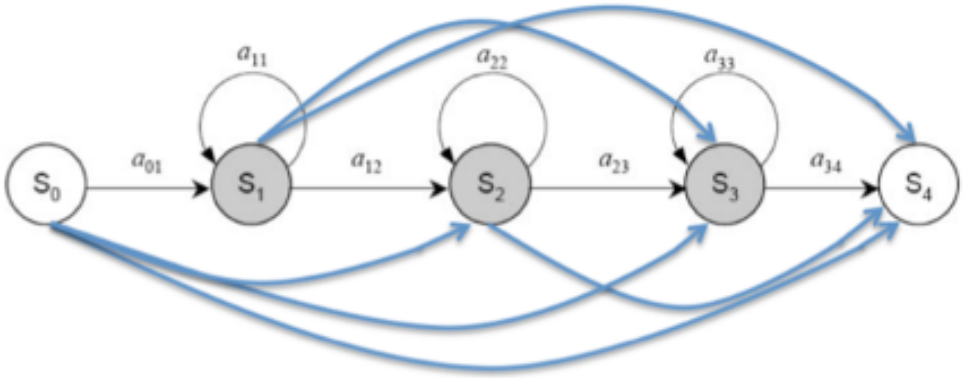

Figure 4.48: HMM with skip-state transitions, reproduced from Yuan, Lai, Cieri, and Liberman (2018, n. p.)

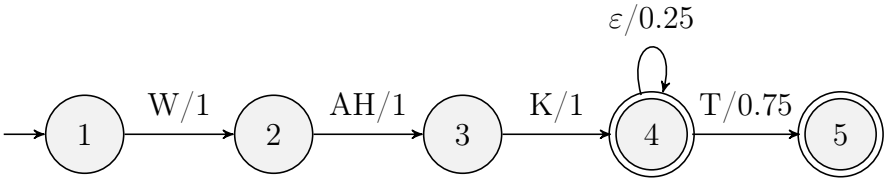

Figure 4.49: Example WFST of *walked*

Chapter 5

Conclusion

5.1 Introduction

In this dissertation, I examined subphonemic variation in terms of phonetic detail resulting from the long-distance coarticulatory properties of stops, phonetic detail consisting of an important acoustic cue for phonological properties (i.e., voicing), and how reduction of stops can be captured using automated methods for large-scale data analysis. This dissertation, overall, has shown that there can be systematic differences in fine phonetic detail that reflect linguistic properties, which is a central claim of proponents of the importance of fine phonetic detail. Chapter 2 shows that there is a relationship between the voicing property of the stop and the fine phonetic detail of the surrounding long-distance acoustics. Chapter 3 clarifies linguistic and social effects on voice onset time, showing that the linguistic variables vowel duration, vowel tenseness, stress, and place of articulation all effect the fine phonetic detail property of VOT as suggested by previous literature. Additionally, this study has shown that there are no positive VOT differences across ethnicity, age, or sex, meaning that if we believe these stops truly are realized differently across these groups, we should focus our efforts on what's happening prior to the burst. Finally, Chapter 4 has shown that the linguistic properties morpheme status, following segment, preceding segment, stress,

frequency, phonological neighborhood density, and voicing were all reflected in the rate of deletion, and that MFA is sensitive to following segment and stress. This dissertation has shown support for the idea that fine phonetic detail varies according to linguistic structure throughout its three body chapters.

The remainder of this chapter will begin with a brief summary of all chapters (section 5.2), presents the overall implications for this research (section 5.3), and finally, concludes (section 5.4).

5.2 Chapter Conclusions

To conclude, I present a brief overview of each chapter of this dissertation. Chapter 1 introduces the concept of phonetic detail / subphonemic variation, a concept central to this entire dissertation, as subphonemic variation arises from coarticulation, differences in acoustic cues, and reduction. In Chapter 2, coarticulation, a process that influences phonetic detail, is discussed, along with voicing patterns beyond "voiced" and "voiceless". The findings of this chapter were that the LPC coefficients of the utterance surrounding the stop varied as the voicing pattern of the stop varied, meaning that there are potentially linguistically-relevant long-distance phonetic detail in the utterance. Chapter 3 examines the acoustic cue voice onset time, another way in which subphonemic properties can vary. No statistically significant sociophonetic variation was found in this acoustic cue. Chapter 4 discusses word-final /t, d/ deletion in consonant clusters, and how this reduced variant may be incorporated into forced alignment technology. Additionally, the subphonemic properties of /t, d/, and how they related to MFA's determination of /t, d/ presence or absence, were examined. The finding of this chapter was that it improved the accuracy of one forced alignment technology, the Montreal Forced Aligner (McAuliffe et al., 2017). An additional finding from the study of subphonemic variation is that the burst influenced MFA's decision of [d] presence, while the closure, burst, and well-attested variants of /t/ influenced MFA's decision of [t] presence.

5.3 Implications

There are five topics discussed below for which the study of phonetic detail is crucial. Two are applications in technology, speech synthesis (subsection 5.3.1) and automatic speech recognition (subsection 5.3.2); two are central to the field of phonetics, speech production (subsection 5.3.3) and speech perception (subsection 5.3.4); and the final topic is exemplar theory (subsection 5.3.5). The five subsections below discuss each of these in turn, in addition to what each chapter's findings mean for each topic.

5.3.1 Speech Synthesis

Speech synthesis (or text-to-speech, TTS) is the process in which a waveform is generated from text. There are two major components to speech synthesis: 1. front-end, in which text is labeled with phonetic transcriptions, "usually includ[ing] phone, syllable, word, phrase, and utterance level features" (Sotelo et al., 2017, p. 1) and 2. back-end, in which these linguistic features are converted into a waveform. This dissertation's line of research, and the areas of research of many other scholars (Schuppler, Ernestus, Van Dommelen, & Koreman, 2010; Schuppler, Van Dommelen, Koreman, & Ernestus, 2009; Warner, Jongman, Sereno, & Kemps, 2004; Warner & Tucker, 2011), to cite only a few, is that qualities beneath the level of the phone are paramount to understanding what makes speech sound natural—its *perceptual coherence*. Sotelo et al. (2017) state that one of the primary goals in speech synthesis today is *naturalness*, which they define as "describ[ing] information not directly captured by intelligibility, such as overall ease of listening, global stylistic consistency, regional or language level nuances, among others" (2017, p. 1). In fact, Hawkins and Smith (2001) argue that naturalness is "possibly the main thing that sets [human speech] apart from most synthetic speech" (2001, p. 104). Even beyond the desire for synthetic speech to sound natural for

aesthetic reasons, decreased naturalness due to lack of phonetic detail in synthetic speech is argued to increase mental processing costs in speech understanding (Duffy & Pisoni, 1992).

This goal of naturalness can be informed by subphonemic detail, such as the kind examined in this dissertation. In Chapter 2, it was determined that the acoustics, in terms of LPC coefficients, vary as the voicing of a stop consonant varies. This reveals subtle differences in acoustics that have implications for speech technology. The findings that the spectral qualities of the waveform differ in alignment with the differing voicing qualities of the stop (and combined further with the fact that phonologically voiced or voiceless stops often have patterns of voicing beyond just "voiced" and "voiceless") must be considered among the linguistic features in text-to-speech front-end processing if the goal of naturalness will ever be achieved. In fact, Hawkins (2003) cites Coleman's (2003) finding of long-distance coarticulation as part of perceptual coherence[1]. Furthermore, regional qualities examined in Chapter 3, such as the subtle differences (or lack thereof) in positive VOT due to sociolinguistic variables can inform models of subphonemic differences. Finally, the subphonemic variation found in Chapter 4 could be incorporated into models in order for synthetic speech to sound more natural.

5.3.2 Automatic Speech Recognition

In automatic speech recognition (ASR), a process in which text is generated from a waveform, the waveform is converted into spectral feature vectors, and each of these vectors is used to derive likelihoods of phones, which is then used to detemine a sequence of most likely words using a dictionary and language model. The success of this system is typically judged on word error rate (WER)[2], and ASR systems have seen a great deal of improvement in recent history, with error rates as low as 5.1% (Xiong et al., 2018). However, speech recognition has been

[1] Additional examples of which are the McGurk Effect (McGurk & MacDonald, 1976) and vowel-to-vowel coarticulation (Alfonso & Baer, 1982).

[2] The shortcomings of using WER as a metric for speech recognition is detailed in Park, Patwardhan, Visweswariah, and Gates (2008), but addressing the potential pitfalls of this equation is outside the scope of this dissertation.

shown to perform worse on female speakers and speakers of certain dialects, e.g., Scottish English (Tatman, 2017). Tatman (2017) found that mean WER for women was approximately 50%, while WER for men was approximately 37%, as shown in Figure 5.1. Furthermore, as shown in Figure 5.2, WER for Scottish English speakers was significantly higher than WER of speakers from other regions. The region with the second highest error rate is Georgia, USA, though the alignment of the DASS (Kretzschmar et al., 2013; Kretzschmar et al., 2019) corpus was found to be fairly accurate, as mentioned in Chapter 3.

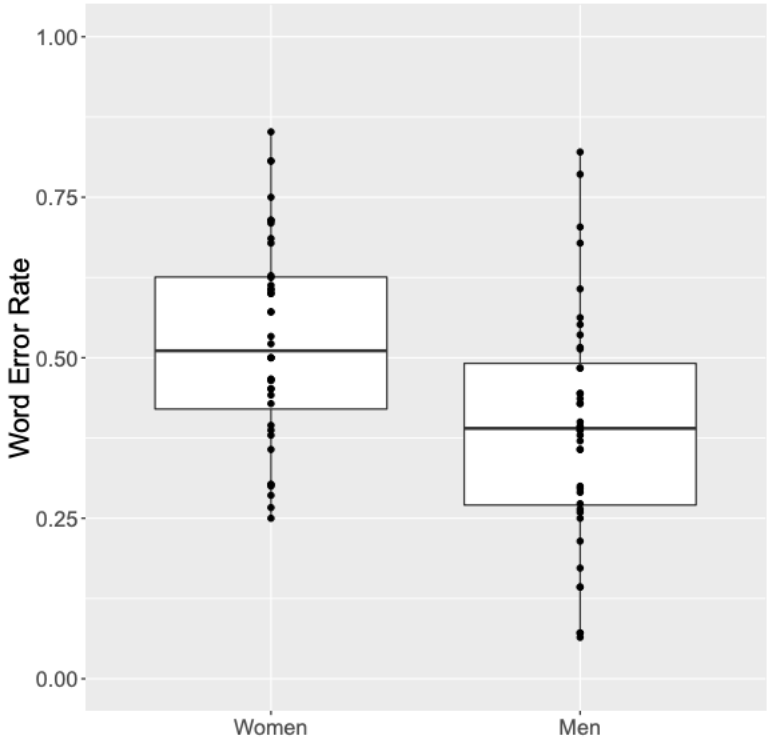

Figure 5.1: Word error rate by sex for speakers of Scottish English, reproduced from Tatman (2017)

In the field of automatic speech recognition, subphonemic variation can make a difference in the system's accuracy. Further understanding the subphonemic differences present in speech, as acknowledged in Chapter 3, can ensure that training data is balanced in order to correct these biases. Forced alignment, an offshoot of automatic speech recognition, can also

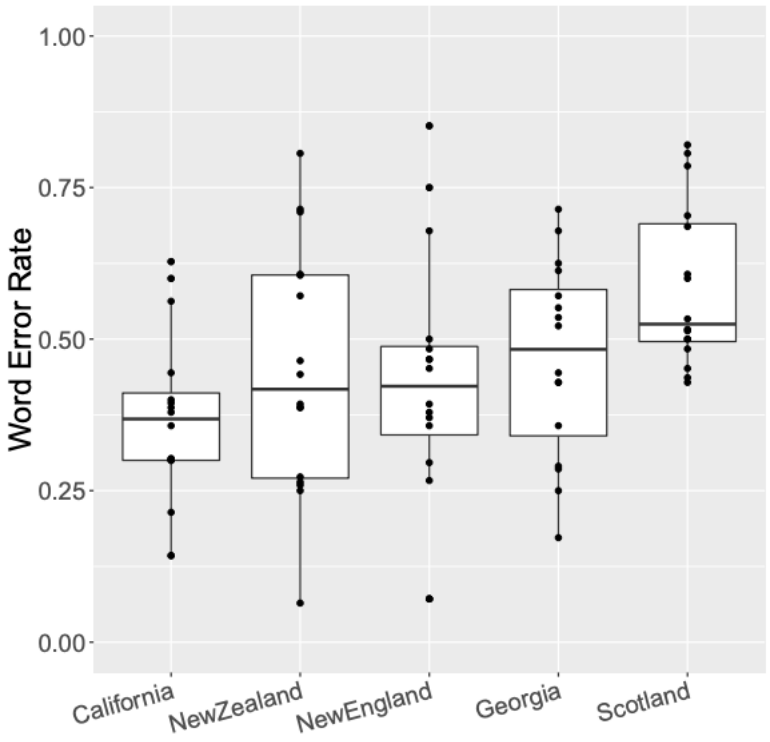

Figure 5.2: Word error rate by dialect, reproduced from Tatman (2017)

be improved with further research on subphonemic properties of speech, which Chapter 4 discusses at length. To summarize briefly part of the findings of Chapter 4 here, dictionary modifications to existing technologies (e.g., MFA) can be made in order to better capture variation during the forced alignment process.

5.3.3 Speech Production

The research regarding subphonemic variation is also important to models of speech production, i.e., how *humans* produce speech. Studies showing subphonemic variation can be captured in probabilistic models, such as Keating's window model (Keating, 1990). Applied

to coarticulation, this model outlines a theory of speech production in which there are certain articulatory "windows" that are necessary for the percept of a certain sound, as shown in the left in Figure 5.3. In Keating's model, the allowable range of coarticulation is indicated with dashed lines. Blackburn and Young (2000) expanded this model to include probability, by using Gaussian distributions to model the fact that not all instances of variation are equally likely. Even further, Coleman et al. (2016) extend this model to include Gaussian mixture models, which account for the fact that not all articulation positions inside these "windows" have one single unimodal Gaussian distributions. The fact that the LPC coefficients can vary with subphonemic variation in voicing (Chapter 2), that VOT shows slight trends in sociolinguistic differences (Chapter 3), and the prevalence of /t, d/ deletion and the differences in the subphonemic properties of /t, d/ (Chapter 4) are all variations that could be modeled probabilistically.

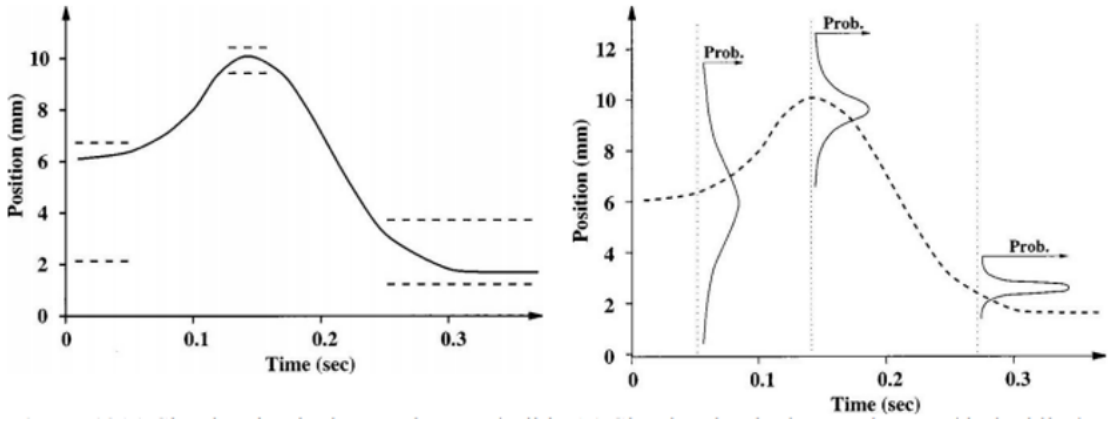

Figure 5.3: Keating's Window Model (left) and Blackburn and Young's model (right), reproduced from Blackburn and Young (2000)

5.3.4 Speech Perception

Finally, this research also has implications for speech perception. Of course, the lack of invariance in speech has been a central question in speech perception studies. For example, researchers have been interested in the intelligibility of conversational speech, which is often

reduced (Pollack & Pickett, 1963). Phonetic reduction, of which deletion is an example, is even argued to help the listener understand what is being said because reduction allows the listener to exploit the knowledge that more frequent words are also more likely to be reduced (Mitterer & Russell, 2013). While there has been research that shows how context (acoustic context (Janse & Ernestus, 2011), semantic context (Van de Ven, Tucker, & Ernestus, 2011), discourse context (Brouwer, Mitterer, & Huettig, 2013), and syntactic context (Viebahn, Ernestus, & McQueen, 2015)) plays a role in top-down theories of speech perception, there is also evidence that acoustic cues, such as VOT, are necessary for perception of reduced speech. For example, Lavoie cites the example of reduced function words: "Each function word seems to have a minimal amount of phonetic material that must be realized for the word to be present" (Lavoie, 2002a, p. 178). The subphonemic analysis in Chapter 4 reveals that for /d/, it is possible that a /d/ must, at minimum, have a burst to be considered present. Furthermore, Van de Ven, Ernestus, and Schreuder (2012) found that when acoustic cues were in conflict with semantic/syntactic cues, listeners relied on subphonemic acoustic information. Clearly, the information on subphonemic processes garnered from Chapter 3 and Chapter 4 have implications for speech processing, and what it means that humans can perceive speech that is not in citation form.

5.3.5 Phonological Theory

In addition to the implications for the subfields mentioned above, these phonetic studies have importance to the phonological theory (i.e., exemplar, abstractionist, hybrid models) discussed in Chapter 1, as well.

The results of Chapter 2 suggest that there is a relationship between word-final voicing patterns and long-distance acoustics. This result is difficult for a segmental or abstractionist approach to phonology to account for, as subtle changes are spread across many segments.

The fact that there is a relationship between production and linguistic structure supports the idea that fine phonetic detail is influenced by linguistic structure.

Further, one would not expect voicing patterns that influence acoustics of the surrounding segments to be random; these are influenced by the linguistic structure. Therefore, it follows that this coarticulation could provide the listener with linguistic knowledge about the utterance. In contrast, for speech perception, one might argue that this variability introduced by the coarticulation is an issue, especially for abstraction. However, this variation is at least somewhat predictable (as evidenced by the results of machine learning), and the variation could be understood through a model that takes linguistic context into account, such as segment position. For example, if you were to compare the phrases *The rainbow* and *The plain bow*, the /b/ of *The rainbow* is word-medial, and more likely to be voiced, while the /b/ of *The plain bow* is word-initial and less likely to be voiced. Subtle differences in the acoustics of these utterances arising from this voicing could then help the listener determine that /b/ of *The plain bow* starts a separate word. This dissertation does not include a perception study, though discrimination of coarticulated tokens could yield relevant results. Following the conclusions of Manker (2020), lack of discriminability in tokens with expected coarticulation as opposed to unexpected phonetic detail would lend support to a hybrid model, where there is "some degree of filtering or abstraction [...] in exemplar storage" (2020, p. 17).

The results of Chapter 3 explore VOT and its high level of variability. This variability could be modeled in a similar way to the coarticulation mentioned above; differences in VOT duration reflect the linguistic variables shown to influence it. Conversely, for speech perception, listeners can exploit the relationship between VOT duration and linguistic features. For example, the word-medial /k/ in "record" would be expected to have a longer VOT in the verb, which has stress on the second syllable, in comparison with the noun, which has stress on the first syllable. This duration of VOT, along with vowel reduction and other factors, can be used to determine stress. Although social variables were not significant in this

dissertation, social factors are known to have consequences for exemplar storage, as Jones and Clopper (2019) provide evidence for Sumner, Kim, King, and McGowan's (2014) claim that "socially idealized forms are more robustly encoded than other forms" (2019, p. 174).

Finally, the results of Chapter 4 regarding the subphonemic properties of word-final /t/ and /d/ show that there are many ways in which these phonemes can vary, and potentially this can be linked to linguistic information or information about the communicative situation. For example, reduction of /t/ to a flap (e.g., *sort of* realized as *sorta*) could indicate a less formal situation. Additionally, several linguistic variables (e.g., morpheme status, stress, etc.) were shown to play a role in deletion, lending support to hybrid models.

Taken together, each of the results mentioned above could be modeled in terms of hybrid abstractionist-exemplar theory of production; "in hybrid exemplar models, production targets *can* be generated on the basis of phonemic categories alone, but will normally be influenced also by larger units" (Smith, Baker, & Hawkins, 2012, p. 50) (e.g., Pierrehumbert (2006)). There are levels of linguistic structure that influence production, but there is potential for some level of abstraction as well.

5.4 Conclusion

This dissertation has examined subphonemic variation in English stops through three studies, and this chapter has endeavored to interpret these results in terms of several areas of interest within the fields of phonetics and phonology: Speech Synthesis, Automatic Speech Recognition, Speech Production, Speech Perception, and Phonological Theory. Overall, my hope is that the reader is convinced that there is interesting variability beneath the level of the phoneme, and that the inclusion of this variation into all areas of concern in phonetics and phonology is paramount to further our understanding of these fields.

Bibliography

Adda-Decker, M., & Lamel, L. (1999). Pronunciation variants across system configuration, language, and speaking style. *Speech Communication*, *29*(2–4), 83–98.

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, *100*(1), 90–93.

Alfonso, P. J., & Baer, T. (1982). Dynamics of vowel articulation. *Language and Speech*, *25*(2), 151–173.

Anderson, S. R. (1985). *Phonology in the twentieth century: Theories of rules and theories of representations*. University of Chicago Press.

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*(3), 349–371.

Avery, P., & Rice, K. (1989). Segment structure and coronal underspecification. *Phonology*, *6*(2), 179–200.

Bagwell, C., & Norskog, L. (2015). SoX — Sound eXchange, the swiss army knife of audio manipulation (Version 14.4.2).

Bailey, G. (2016). Automatic detection of sociolinguistic variation using forced alignment. In *University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV 44)* (pp. 10–20). York.

Bailey, G., Wikle, T., Tillery, J., & Sand, L. (1991). The apparent time construct. *Language Variation and Change*, *3*(3), 241–264.

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, *52*(4), 391–413.

Baranowski, M. (2013). Sociophonetics. In R. Bayley, R. Cameron, & C. Lucas (Eds.), *The Oxford Handbook of Sociolinguistics*. Oxford University Press.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., ... Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation* (pp. 1–55). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.wmt-1.1

Barry, M. (1985). A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics*, *4*, 1–16.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bayley, R. (1994). Consonant cluster reduction in Tejano English. *Language Variation and Change*, *6*(3), 303–326.

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259–284.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Berry, J. (2004). Control of short lag VOT (voice-onset time) for voiced English stops. *The Journal of the Acoustical Society of America*, *115*(5), 2465.

Bigi, B., & Hirst, D. (2012). SPeech Phonetization Alignment and Syllabification (SPPAS): A tool for the automatic analysis of speech prosody. *The Eighth International Conference on Language Resources and Evaluation*, 19–22. Retrieved from http://www.sppas.org

Blackburn, C. S., & Young, S. (2000). A self-learning predictive model of articulator movements during speech production. *Journal of the Acoustical Society of America*, *107*(3), 1659–1670.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer (Version 6.0.42).

Bogert, B. P. (1963). The quefrency alanysis [sic] of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Time Series Analysis*, 209–243.

Brouwer, S., Mitterer, H., & Huettig, F. (2013). Discourse context and the recognition of reduced and canonical spoken words. *Applied Psycholinguistics*, *34*(3), 519–539.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Butcher, A., & Weiher, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics*, *4*(1), 59–74.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261–290.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, *15*(1–2), 39–54.

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 629–647.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *32*(2), 141–176.

Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics*, *64*, 71–89.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47.

Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, *48*(6), 633–644.

Clopper, C. G., Turnbull, R., Cangemi, F., Clayards, M., Niebuhr, O., Schuppler, B., & Zellers, M. (2018). Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. *Rethinking Reduction*, 25–72.

Coetzee, A. W., Beddor, P. S., Styler, W., Tobin, S., Bekker, I., & Wissing, D. (2019). Producing and perceiving socially indexed coarticulation in Afrikaans. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 215–219).

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Cohen Priva, U., & Gleason, E. (2020). The causal structure of lenition: A case for the causal precedence of durational shortening. *Language*, *96*(2), 413–448.

Cohen Priva, U., & Jurafsky, D. (2008). *Phone information content influences phone duration.* A poster presented at Prosody08, Cornell University.

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

Cohen, P., & Labov, W. (1963). Systematic relations of standard and non-standard rules in the grammars of Negro Speakers. *Language, Society, and Education: A profile of Black English*, 149–160.

Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics*, *31*(3), 351–372.

Coleman, J. (2017). Why should linguists study acoustic phonetics? Retrieved from http://www.phon.ox.ac.uk/jcoleman/why_acoustics.htm

Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012). Audio BNC: The audio edition of the Spoken British National Corpus. *Phonetics Laboratory, University of Oxford*. Retrieved from http://www.phon.ox.ac.uk/AudioBNC

Coleman, J., Renwick, M. E., & Temple, R. (2016). Probabilistic phonetic underspecification in nasal place assimilation. *Phonology*, *33*(3), 425–458.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology*, *10*, 91–111.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). xtable: Export tables to LaTeXor HTML (Version R package version 1.8-4). Retrieved from https://CRAN.R-project.org/package=xtable

Dall, R., Brognaux, S., Richmond, K., Valentini-Botinhao, C., Henter, G. E., Hirschberg, J., ... King, S. (2016). Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5155–5159). Shanghai: IEEE.

Daniloff, R. G., & Hammarberg, R. E. (1973). On defining coarticulation. *Journal of Phonetics*, *1*(3), 239–248.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, *54*, 35–50.

Davidson, L. (2018). Phonation and laryngeal specification in American English voiceless obstruents. *Journal of the International Phonetic Association*, *48*(3), 331–356.

Davis, F., & Cohn, A. C. (2020). The relationship between lexical frequency, compositionality, and phonological reduction in English compounds. In *94th Annual Meeting of the Linguistic Society of America*.

De Jong, K., Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech*, *36*(2–3), 197–212.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*(4), 769–773.

Docherty, G. J., & Ladd, D. R. (1992). *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press.

Docherty, G. J., Watt, D., Llamas, C., Hall, D., & Nycz, J. (2011). Variation in voice onset time along the Scottish-English border. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 591–594).

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, *35*(4), 351–389.

Eager, C. (2015). Automated voicing analysis in Praat: Statistically equivalent to manual segmentation. *The Scottish Consortium for International Congress of Phonetic Sciences*, *18*, 551–585.

Eddington, D., Treiman, R., & Elzinga, D. (2013). Syllabification of American English: Evidence from a large-scale experiment. Part I. *Journal of Quantitative Linguistics*, *20*(1), 45–67.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface*. Utrecht: LOT.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, *142*, 27–41.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, *39*(3), 253–260.

Esplà-Gomis, M., Forcada, M. L., Ramírez-Sánchez, G., & Hoang, H. (2019). Paracrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks* (pp. 118–119).

Fairbanks, G. (1940). *Voice and Articulation Drillbook*. New York: Harper.

Fasold, R. W. (1972). Tense marking in Black English: A linguistic and social analysis. In R. W. Shuy (Ed.), *Urban Language Series, No. 8*. Arlington, Virginia: Center for Applied Linguistics.

Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word*, *14*(1), 47–56.

Flege, J. E. (2007). Language contact in bilingualism: Phonetic system interactions. *Laboratory Phonology*, *9*, 353–382.

Flege, J. E., Frieda, E. M., Walley, A. C., & Randazza, L. A. (1998). Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, *20*(2), 155–187.

Flege, J. E., & Brown Jr., W. S. (1982). The voicing contrast between English /p/ and /b/ as a function of stress and position-in-utterance. *Journal of Phonetics*, *10*(4), 335–345.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *The Journal of the Acoustical Society of America*, *138*(4), 2132–2139.

Forrest, J. (2017). The dynamic interaction between lexical and contextual frequency: A case study of (ing). *Language Variation and Change*, *29*(2), 129–156.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, *34*(4), 409–438.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*(5), 489–504.

Frank, E., Hall, M. A., & Pal, C. J. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, *27*(4), 765–768.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, *31*(7), 1–24. doi:10.18637/jss.v031.i07

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251.

Gonet, W., & Święciński, R. (2012). More on the voicing of English obstruents: Voicing retention vs. voicing loss. *Research in Language*, *10*(2), 183–199.

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, *1*.

Gordon, M. J. (2005). Do You Speak American? Retrieved January 20, 2020, from https://www.pbs.org/speak/ahead/change/changin/

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 163–179.

Guy, G. (1980). Variation in the group and the individual: The case of final stop deletion. In *Locating Language in Time and Space* (pp. 1–36). Academic Press.

Guy, G. R. (2014). Linking usage and grammar: Generative phonology, exemplar theory, and variable rules. *Lingua*, *142*, 57–65.

Guy, G. R., Hay, J., & Walker, A. (2008). Phonological, lexical, and frequency factors in coronal stop deletion in early New Zealand English. *Laboratory Phonology*, *11*, 53–54.

Halle, M., & Stevens, K. N. (1971). A note on laryngeal features. *Quarterly Progress Report 101*, 198–212.

Hawkins, S. (2003). Contribution of fine phonetic detail to speech understanding. In *Proceedings of the 15th International Congress of Phonetic Sciences* (Vol. 293, p. 296).

Hawkins, S. (2010). Phonetic variation as communicative system: Perception of the particular and the abstract. *Laboratory Phonology*, *10*, 479–510.

Hawkins, S. (2012). The lexicon: Not just elusive, but illusory? In A. C. Cohn, C. Fougeron, & M. K. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 162–173). Oxford: Oxford University Press.

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics*, *13*, 99–188.

Hayes, B. (2011). *Introductory phonology*. John Wiley & Sons.

Heid, S., & Hawkins, S. (2000). An acoustical study of long-domain /r/ and /l/ coarticulation. In *Proceedings of the 5th Seminar on Speech Production: Models and Data* (pp. 77–80). Kloster Seeon Germany.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738–1752.

Hoit, J. D., Solomon, N. P., & Hixon, T. J. (1993). Effect of lung volume on voice onset time (VOT). *Journal of Speech, Language, and Hearing Research*, *36*(3), 516–520.

Honeybone, P. (2005). Diachronic evidence in segmental phonology: The case of obstruent laryngeal specifications. *The Internal Organization of Phonological Segments*, 319–354.

Hoole, P., Nguyen-Trong, N., & Hardcastle, W. (1993). A comparative investigation of coarticulation in fricatives: Electropalatographic, electromagnetic, and acoustic data. *Language and Speech*, *36*(2–3), 235–260.

Hooper, J. B. (1976). Word frequency in lexical diffusion and the source of morphophonological change. *Current Progress in Historical Linguistics*, *96*, 105.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, *24*(2), 225–232.

Hunnicutt, L., & Morris, P. A. (2016). Prevoicing and Aspiration in Southern American English. *University of Pennsylvania Working Papers in Linguistics*, *22*(1), 215–224.

Husain, T. M. (2017). Acoustic measurement of voiced implosives: Evidence of voiced implosives in a US dialect. *Southern Journal of Linguistics*, *41*(1), 62–87.

Irfana, M., & Sreedevi, N. (2019). Extent of Lingual Coarticulation: A Crosslinguistic Study Using Ultrasound Imaging. *Research & Reviews: Journal of Medical Science and Technology*, *8*(3), 22–31.

Jacewicz, E., Fox, R. A., & Lyle, S. (2009). Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association*, *39*(3), 313–334.

Jakobson, R., & Waugh, L. R. (2002). *The sound shape of language*. Walter de Gruyter.

Janse, E., & Ernestus, M. (2011). The roles of bottom-up and top-down information in the recognition of reduced speech: Evidence from listeners with normal and impaired hearing. *Journal of Phonetics*, *39*(3), 330–343.

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

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis. Proceedings of the 1st Session of the 10th International Symposium* (pp. 29–54).

Johnson, K. (2012). *Acoustic and Auditory Phonetics* (3rd ed.). Malden: Wiley-Blackwell.

Jones, D. (1956). *An Outline of English Phonetics* (8th ed.). Cambridge University Press.

Jones, J. A., & Renwick, M. E. (2020). Spatial analysis of sub-regional variation in Southern US English. *Journal of Linguistic Geography*, To appear.

Jones, Z., & Clopper, C. G. (2019). Subphonemic variation and lexical processing: Social and stylistic factors. *Phonetica*, *76*(2–3), 163–178.

Jurafsky, D., & Martin, J. H. (2000). *Speech & Language Processing* (1st ed.). Pearson Education India.

Jurafsky, D., & Martin, J. H. (2008). *Speech & Language Processing* (2nd ed.). Pearson Education India.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. (1998). Reduction of English function words in switchboard. In *Fifth International Conference on Spoken Language Processing*.

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of the lemma in form variation. *Laboratory Phonology*, *7*.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). *The effect of language model probability on pronunciation reduction.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2001, Salt Lake City, UT.

Jurafsky, D., & Martin, J. H. (2019). Speech & language processing (draft). Retrieved from https://web.%20stanford.%20edu/~jurafsky/slp3

Keating, P. (1998). Word-level phonetic variation in large speech corpora. *ZAS Papers in Linguistics*, *11*, 35–50.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, 286–319.

Keating, P. A. (1990). The window model of coarticulation: Articulatory evidence. *Papers in Laboratory Phonology I*, *26*, 451–470.

Keating, P., & Lahiri, A. (1993). Fronted velars, palatalized velars, and palatals. *Phonetica*, *50*(2), 73–101.

Kelly, J., & Local, J. (1989). Doing phonology: Observing, recording, interpreting. *England: Manchester University Press*.

Keshet, J., Sonderegger, M., & Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction (Version 0.91).

Kiesling, S., Dilley, L., & Raymond, W. D. (2006). The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of Conversational Speech.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual Processing of Speech via Web Services. *Computer Speech & Language*, *45*, 326–347.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, *18*(4), 686–706.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Klatt, D. H., & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In *Structure and Process in Speech Perception* (pp. 69–89). Springer.

Kochetov, A., & Neufeld, C. (2013). Examining the extent of anticipatory coronal coarticulation: A long-term average spectrum analysis. In *Proceedings of meetings on acoustics* (Vol. 19).

Kretzschmar, W. A., Bounds, P., Hettel, J., Pederson, L., Juuso, I., Opas-Hänninen, L. L., & Seppänen, T. (2013). The Digital Archive of Southern Speech (DASS). *Southern Journal of Linguistics*, *37*(2), 17–38.

Kretzschmar, W. A. J., Renwick, M. E., Lipani, L., Olsen, M., Olsen, R. M., Shi, Y., & Stanley, J. A. (2019). *Transcriptions of the Digital Archive of Southern Speech.* Linguistic Atlas Project, University of Georgia. Retrieved from http://www.lap.uga.edu/Projects/DASS2019/

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2016). lmerTest: Tests in Linear Mixed Effects Models. (Version R package version 2.0-32). Retrieved from https://CRAN.R-project.org/package=lmerTest

Labov, W. (1968). *A study of the non-standard English of Negro and Puerto Rican speakers in New York City.* Columbia University.

Labov, W. (1972). *Sociolinguistic Patterns.* University of Pennsylvania Press.

Labov, W. (1986). The social stratification of (r) in new york city department stores. In *Dialect and language variation* (pp. 304–329). Elsevier.

Labov, W. (2006). *The Social Stratification of English in New York City.* Cambridge University Press.

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English.* Berlin: Mouton de Gruyter.

Labov, W., Yaeger, M., & Steiner, R. (1972). *A Quantitative Study of Sound Change in Progress.* US Regional Survey.

Ladefoged, P., & Johnson, K. (2014). *A Course in Phonetics.* Nelson Education.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. *Laboratory phonology*, *7*, 637–675.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*(1), 44–59.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Lavoie, L. (2002a). Some influences on the realization of for and four in American English. *Journal of the International Phonetic Association*, *32*(2), 175–202.

Lavoie, L. (2002b). Subphonemic and suballophonic consonant variation: The role of the phoneme inventory. *ZAS papers in linguistics*, *28*, 39–54.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Liberman, P., & Blumstein, S. (1988). Speech Physiology. *Speech Perception, and Acoustic*.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In *Speech Production and Speech Modelling* (pp. 403–439). Dordrecht: Springer.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

Lipani, L. (2017). Word-final velar place assimilation in English. *Proceedings of the Linguistic Society of America*, *2*(24), 1–15.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, *10*(1), 1–28.

Lison, P., & J., T. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Local, J. (2003). Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics*, *31*, 321–339.

Local, J. (2007). Phonetic detail and the organisation of talk-in-interaction. *Proceedings of the 16th ICPhS, Saarbrücken, Germany*.

Luczak, M. (2018). Combining raw and normalized data in multivariate time series classification with dynamic time warping. *Journal of Intelligent & Fuzzy Systems*, *34*(1), 373–380.

Manker, J. (2020). The perceptual filtering of predictable coarticulation in exemplar memory. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1), 1–17.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech* (pp. 498–502).

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., ... Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Milne, P. (2014). *The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French* (Doctoral dissertation, University of Ottawa Ottawa, Ontario, Canada).

Mitterer, H. (2008). How are words reduced in spontaneous speech? In *ISCA Workshop on Experimental Linguistics 2008* (pp. 165–168). University of Athens.

Mitterer, H. (2011). Recognizing reduced forms: Different processing mechanisms for similar reductions. *Journal of Phonetics*, *39*(3), 298–303.

Mitterer, H., & Russell, K. (2013). How phonological reductions sometimes help the listener. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 977.

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, *16*(1), 69–88.

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, *36*(2), 308–317.

Mueen, A., & Keogh, K. (2016). *Extracting Optimal Performance from Dynamic Time Warping*. Retrieved from https://www.cs.unm.edu/~mueen/DTW.pdf

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, *47*, 1048–1058.

Nagy, N., & Kochetov, A. (2013). Voice onset time across the generations: A cross-linguistic study of contact-induced change, 19–38.

Newlin-Łukowicz, L. (2014). From interference to transfer in language contact: Variation in voice onset time. *Language Variation and Change*, *26*(3), 359–385.

Nguyen, N., Wauquier, S., & Tuller, B. (2009). The dynamical approach to speech perception: From fine phonetic detail to abstract phonological categories. *Approaches to Phonological Complexity*, 193–217.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization*, 18–39.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168.

Olive, J. P., Greenwood, A., & Coleman, J. (1993). *Acoustics of American English speech: A dynamic approach*. Springer Science & Business Media.

Olsen, R. M., Olsen, M. L., & Renwick, M. E. (2017). The impact of sub-region on /aɪ/ weakening in the US South. In *Proceedings of Meetings on Acoustics 174ASA* (Vol. 31), ASA.

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 (ICASSP)* (pp. 5206–5210).

Park, Y., Patwardhan, S., Visweswariah, K., & Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. In *Ninth Annual Conference of the International Speech Communication Association*.

Pederson, L., McDaniel, S. L., & Adams, C. M. (Eds.). (1986). *Linguistics Atlas of the Gulf States*. Athens, Georgia: University of Georgia Press.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency. *Frequency and the Emergence of Linguistic Structure*, *45*, 137–157.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *4*(34), 516–530.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, *2*, 33–52.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569.

Pollack, I., & Pickett, J. M. (1963). The intelligibility of excerpts from conversation. *Language and Speech*, *6*(3), 165–171.

Porter, M. F. et al. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Technical Report DAMTP 2009/NA06*.

Quinian, J. R. (1993). *C4.5: Programs for Machine Learning.*

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.1). Vienna, Austria. Retrieved from http://www.R-project.org/

Rao, K. S., & Manjunath, K. (2017). *Speech recognition using articulatory and excitation source features.* Springer.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303.

Raymond, W. D., Brown, E. L., & Healy, A. F. (2016). Cumulative context effects and variant lexical representations: Word use and English final t/d deletion. *Language Variation and Change*, *28*(2), 175–202.

Raymond, W. D., Dautricourt, R., & Hume, E. (2006). Word-internal/t, d/deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, *18*(1), 55–97.

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

Read, C. (1971). Pre-school children's knowledge of English phonology. *Harvard Educational Review*, *41*(1), 1–34.

Recasens, D. (2018, February 26). Coarticulation. *Oxford Research Encyclopedia of Linguistics.* doi:10.1093/acrefore/9780199384655.013.416

Reetz, H., & Jongman, A. (2011). *Phonetics: Transcription, Production, Acoustics, and Perception.* John Wiley & Sons.

Renwick, M. E. L., Baghai-Ravary, L., Temple, R., & Coleman, J. (2014). *(t,d) deletion in everyday speech.* Poster presented at LabPhon 14, the 14th Conference on Laboratory Phonology, Tokyo, Japan.

Renwick, M. E., & Olsen, R. M. (2015). Voices of coastal Georgia. In *Proceedings of Meetings on Acoustics 170ASA* (Vol. 25), ASA.

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

Renwick, M. E., & Stanley, J. A. (2017). Static and dynamic approaches to vowel shifting in the Digital Archive of Southern Speech. In *Proceedings of Meetings on Acoustics 173EAA* (Vol. 30, *1*), ASA.

Riebold, J. M. (2011). Time to pull out the stops: Spirantization in Pacific Northwestern English. *The Journal of the Acoustical Society of America*, *129*(4), 2453–2453.

Robinson, D. (2019). *Fuzzyjoin: Join tables together on inexact matching.* R package version 0.1.5. Retrieved from https://CRAN.R-project.org/package=fuzzyjoin

Rogers, H. (2014). *The Sounds of Language: An introduction to phonetics.* London: Routledge.

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. Retrieved from http://fave.ling.upenn.edu

RStudio. (2017). RStudio: Integrated development environment for R (Version 1.0.153). Boston, MA. Retrieved from http://www.rstudio.org/

Ryalls, J., Simon, M., & Thomason, J. (2004). Voice onset time production in older Caucasian- and African-Americans. *Journal of Multilingual Communication Disorders*, *2*(1), 61–67.

Ryalls, J., Zipprer, A., & Baldauff, P. (1997). A preliminary investigation of the effects of gender and race on voice onset time. *Journal of Speech, Language, and Hearing Research*, *40*(3), 642–645.

Santa Ana, O. (1991). *Phonetic simplification processes in English of the Barrio: A cross-generational sociolinguistic study of the Chicanos of Los Angeles* (Doctoral dissertation, University of Pennsylvania).

Schuppler, B. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*, 96–109.

Schuppler, B. (2012). How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. *Journal of Phonetics*, *40*, 595–607.

Schuppler, B., Ernestus, M., Van Dommelen, W., & Koreman, J. (2010). Predicting human perception and ASR classification of word-final [t] by its acoustic sub-segmental properties. In *Proceedings of Interspeech 2010* (pp. 2466–2469).

Schuppler, B., Van Dommelen, W., Koreman, J. C., & Ernestus, M. (2009). Word-final [t]-deletion: An analysis on the segmental and sub-segmental level. In *Proceedings of Interspeech 2009* (pp. 2275–2278).

Scobbie, J. M. (2005). Flexibility in the face of incompatible English VOT systems. Berlin: Mouton de Gruyter.

Seyfarth, S., & Garellek, M. (2020). Physical and phonological causes of coda/t/glottalization in the mainstream American English of central Ohio. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *11*(1).

Sharoff, S. (2016). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus*. Retrieved from http://wackybook.sslmit.unibo.it/

Shi, Y. (2019). *An investigation on prenasal merger in Southern American English through automatic speech recognition* (Master's thesis, University of Georgia, Athens, GA).

Shockey, L. (2008). *Sound patterns of spoken English.* John Wiley & Sons.

Shrem, Y., Goldrick, M., & Keshet, J. (2019). Dr. VOT: Measuring Positive and Negative Voice Onset Time in the Wild. *Proceedings of Interspeech 2019*, 629–633.

Simpson, A. P. (2012). The first and second harmonics should not be used to measure breathiness in male and female voices. *Journal of Phonetics*, *40*(3), 477–490.

Smith, B. J., Mielke, J., Magloughlin, L., & Wilbanks, E. (2019). Sound change and coarticulatory variability involving English /r/. *Glossa: A journal of general linguistics*, *4*(1).

Smith, B. L. (1978). Effects of place of articulation and vowel environment on voiced stop consonant production. *Glossa*, *12*(2), 163–175.

Smith, R., Baker, R., & Hawkins, S. (2012). Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *Journal of Phonetics*, *40*(5), 689–705.

Smolensky, P., & Prince, A. (2004). Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in Phonology*, *3*.

Solé, M.-J. (2018). Articulatory adjustments in initial voiced stops in Spanish, French and English. *Journal of Phonetics*, *66*, 217–241.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y. (2017). Char2wav: End-to-end speech synthesis.

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018). Luminosoinsight/wordfreq: V2.2.

Stanley, J. A., Kretzschmar, W. A., Renwick, M. E. L., Olsen, M. L., & Olsen, R. M. (2017). *Gazetteer of Southern Vowels*.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In P. B. Dennes & E. E. David Jr. (Eds.), *Human Communication, A Unified View*. McGraw Hill.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. *Perspectives on the Study of Speech*, 1–38.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329–353.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, *6*(3–4), 505–549.

Styler, W. (2013). Using Praat for linguistic research. *University of Colorado at Boulder Phonetics Lab*.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, *4*, 1015.

Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, *75*(3), 983–992.

Tabain, M. (2019). An Electropalatographic Study of Variability in Arrernte Consonant Production. *Phonetica*, 1–30.

Tagliamonte, S., & Temple, R. (2005). New perspectives on an ol' variable: (t, d) in British English. *Language Variation and Change*, *17*(3), 281–302.

Tamminga, M. (2018). Modulation of the following segment effect on English coronal stop deletion by syntactic boundaries. *Glossa: A Journal of General Linguistics*, *3*(1).

Tamminga, M., & Fruehwald, J. (2013). *Deconstructing TD deletion*. Paper presented at NWAV. University of Pittsburgh/Carnegie Mellon University.

Tanner, J., Sonderegger, M., & Wagner, M. (2017). Production planning and coronal stop deletion in spontaneous speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1).

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the first ACL Workshop on Ethics in Natural Language Processing* (pp. 53–59).

Temple, R. (2014). Where and what is (t, d)?: A case study in taking a step back in order to advance sociophonetics. In *Advances in Sociophonetics* (pp. 97–136). Amsterdam: John Benjamins Publishing Company.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982.

Thomas, E. (2001). *An Acoustic Analysis of Vowel Variation in New World English*. Duke University Press.

Thomas, E. (2016). Sociophonetics of consonantal variation. *Annual Review of Linguistics*, *2*, 95–113.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC Press. Retrieved from http://www.dcc.fc.up.pt/~lorgo/DataMiningWithR

Tunley, A. (1999). *Coarticulatory influences of liquids on vowels in English* (Doctoral dissertation, University of Cambridge).

Turnbull, R. (2018). Patterns of probabilistic segment deletion/reduction in English and Japanese. *Linguistics Vanguard*, *4*(s2).

Tuszynski, J. (2019). caTools: Tools: Moving window statistics, GIF, Base64, ROC AUC, etc. (Version R package version 1.17.1.2). Retrieved from https://CRAN.R-project.org/package=caTools

Urbanek, S. (2019). rJava: Low-Level R to Java Interface (Version R package version 0.9-11). Retrieved from https://CRAN.R-project.org/package=rJava

Van de Ven, M., Ernestus, M., & Schreuder, R. (2012). Predicting acoustically reduced words in spontaneous speech: The role of semantic/syntactic and acoustic cues in context. *Laboratory Phonology*, *3*(2), 455–481.

Van de Ven, M., Tucker, B. V., & Ernestus, M. (2011). Semantic context effects in the comprehension of reduced pronunciation variants. *Memory & Cognition*, *39*(7), 1301–1316.

van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111–122. Retrieved from https://CRAN.R-project.org/package=stringdist

Vasilescu, I., Chitoran, I., Vieru, B., Adda-Decker, M., Candea, M., Lamel, L., & Niculescu, O. (2019). Studying variation in Romanian: Deletion of the definite article -l in continuous speech. *Linguistics Vanguard*, *5*(1).

Viebahn, M., Ernestus, M., & McQueen, J. M. (2015). Syntactic Predictability in the Recognition of Carefully and Casually Produced Speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1684–1702.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, *13*(2), 260–269.

Volz, K. G., & von Cramon, D. Y. (2006). What neuroscience can tell about intuitive processes in the context of perceptual discovery. *Journal of Cognitive Neuroscience*, *18*(12), 2077–2087.

Walker, A. (2019). *Using performance to explore consonant voicing in Southern US English.* Paper presented at the Southeastern Conference on Linguistics. Boca Raton, FL. May 30 – June 2.

Walker, A. (2020). Voiced stops in the command performance of Southern US English. *The Journal of the Acoustical Society of America*, *147*(1), 606–615.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, *32*(2), 251–276.

Warner, N., & Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *The Journal of the Acoustical Society of America*, *130*(3), 1606–1617.

Weide, R. L. (1998). Carnegie Mellon Pronouncing Dictionary (Version release 0.6). Retrieved from www.cs.cmu.edu

Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics*, *7*(2), 197–204.

West, P. (1999). The extent of coarticulation of English liquids: An acoustic and articulatory study. In *Proceedings of the International Conference of Phonetic Sciences* (pp. 1901–1904).

Westbury, J. R. (1979). Aspects of the temporal control of voicing in consonant clusters in English. In *Texas Linguistic Forum 14*, Department of Linguistics, University of Texas at Austin.

Whiteside, S. P., & Irving, C. J. (1997). Speakers' sex differences in voice onset time: Some preliminary findings. *Perceptual and motor skills*, *85*(2), 459–463E.

Whiteside, S. P., & Irving, C. J. (1998). Speakers' sex differences in voice onset time: A study of isolated word production. *Perceptual and Motor Skills*, *86*(2), 651–654.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2017). Tidyverse: Easily install and load the 'Tidyverse' (Version 1.2.1). Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., & Francois, R. (2015). dplyr: A grammar of data manipulation. (Version 0.4.3). Retrieved from http://CRAN.R-project.org/package=dplyr

Wolfram, W. A. (1969). *A sociolinguistic description of Detroit Negro Speech.* Arlington, Virginia: Center for Applied Linguistics.

Wuertz, D., Setz, T., & Chalabi, Y. (2017). timeSeries: Rmetrics — Financial Time Series Objects (Version 3042.102). Retrieved from https://CRAN.R-project.org/package=timeSeries

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5934–5938).

Xuanda, C., Ziyu, X., & Jian, H. (2018). The trajectory of voice onset time with vocal aging. *Interspeech Conference 2018*, 1556–1560.

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. *UC Berkeley PhonLab Annual Report*, *5*(5), 29–43.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... Povey, D. (2002). *The HTK book*. Cambridge University Engineering Department.

Yuan, J., Lai, W., Cieri, C., & Liberman, M. (2018). Using forced alignment for phonetics research. *Chinese Language Resources and Processing: Text, Speech and Language Technology.*

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878–3882.

Yuan, J., & Liberman, M. (2011). Automatic detection of "g-dropping" in American English using forced alignment. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 490–493).

Yuan, J., Lin, H., & Liu, Y. (2020). Detection and analysis of t/d deletion in Librispeech. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7324–7328).

Zellou, G., & Tamminga, M. (2014). Nasal coarticulation changes over time in Philadelphia English. *Journal of Phonetics*, *47*, 18–35.

Appendix A

Sentences in the Rainbow Passage

1. When sunlight strikes the raindrops in the air, they act like a prism and form a rainbow.
2. The rainbow is a division of white light into many beautiful colors.
3. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon.
4. There is, according to legend, a boiling pot of gold at one end.
5. People look, but no one ever finds it.
6. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Appendix B

Stop Words in DASS

- able
- about
- across
- after
- all
- almost
- also
- am
- among
- an
- and
- any
- are
- as
- at
- be
- because
- been
- but
- by
- can
- cannot
- could
- dear
- did
- do
- does
- either
- else
- ever
- every
- for
- from
- get
- got
- had
- has
- have
- he
- her
- hers
- him
- his
- how
- however
- i
- if
- in
- into
- is
- it
- its
- just
- least
- let
- like
- likely
- may
- me
- might
- most
- must
- my
- neither

- no
- nor
- not
- of
- off
- often
- on
- only
- or
- other
- our
- own
- rather
- said
- say
- says
- she
- should
- since
- so
- some
- than
- that
- the
- their
- them
- then
- there
- these
- they
- this
- tis
- to
- too
- twas
- us
- wants
- was
- we
- were
- what
- when
- where
- which
- while
- who
- whom
- why
- will
- with
- would
- yet
- you
- your