Structured phonetic variation across dialects and speakers of English and Japanese

James Tanner

Department of Linguistics

McGill University, Montréal

May 2020

A thesis submitted to McGill University in partial fulfillment for the requirements of the degree of Doctor of Philosophy

Abstract

Phonetic variation within languages – across dialects and speakers – has been of longstanding interest to researchers in phonetics and sociolinguistics, as understanding the structures and sources of within-language variability is essential for addressing a range of questions which are core to understanding language. Much research on language-internal variation has focused on studies of single communities restricted by practical constraints; cross-dialectal work has been largely constrained to research on vowel quality, leaving unanswered a range of empirical questions regarding language-internal variation. A recent turn in linguistic research, towards the use of large datasets of multiple speech communities made possible by changes in data access and advances in signal processing tools and statistical modelling, has provided the possibility to both address new questions and re-assess current theoretical perspectives.

This dissertation applies this 'large-scale' approach to the study of structured phonetic variation across dialects and individual speakers of two languages: English and Japanese. Specifically, this research demonstrates how variation may be constrained across multiple dimensions, pointing to ways in which phonetic variation may be structured across dialects and speakers in systematic ways.

Study 1 examines how speakers of Japanese vary in the realisation of the stop voicing contrast, particularly in the use of stop aspiration and closure voicing, using a large corpus of spontaneous speech. Through statistical modelling it was found that, in spite of variation in the overall use of cues across speakers, speakers showed strong relationships in the use of individual cues to mark the voicing contrasts. These relationships were weaker across cues compared with previous work on English and German, suggesting that the structure of this variation across speakers is language-specific, where the underlying specification of phonological contrasts constrains the dimensions of phonetic variability.

The next two studies shift in language to English by utilising data and methods as part of the SPeech Across Dialects of English project. Study 2 investigates dialectal and speaker variation in the English voicing effect – the difference in vowel duration before voiced and voiceless consonants – examined by integrating data from 15 corpora (30 dialects). The results demonstrated that the size of the voicing effect was smaller in spontaneous speech compared with laboratory speech. It was also observed that English exhibits a wide range of sizes across dialects, whilst speakers vary little from their dialectal baselines. These findings suggest that the voicing effect is both more subtly-controlled and more variable than previously reported, whilst remaining remarkably stable within speech communities.

Study 3 further applies the multi-corpus approach to examine dialectal variation in English vowels using data from 11 corpora (21 dialects). This study considered how multiple properties of vowels – their position in formant space, the shape of the formant trajectory, and duration – can characterise the principal dimensions of variability across dialects of English. Through the application of both classification and dimensionality reduction, it was found that all measures were highly informative in defining how vowels vary across English dialects. The relative role of each measure is highly vowel- and dialect-specific, indicating that some vowels are better characterised by some acoustic properties than others.

Together, these findings demonstrate the utility and role of 'large-scale' studies in addressing central questions about the study of phonetic variation. The use of multiple speech corpora and the application of statistical techniques to model patterns of interest in unbalanced data make it increasingly possible to reveal the extent of phonetic variability apparent at multiple levels of linguistic structure.

Resumé

La recherche en phonétique et en sociolinguistique s'intéresse depuis longtemps à la variation qui existe entre les dialectes et les locuteurs car comprendre les structures et les sources de la variabilité est essentiel pour la compréhension du langage. De nombreuses études sur la variation se sont concentrées sur des communautés linguistiques uniques et ont été limitées par des contraintes pratiques; les travaux inter-dialectaux systématiques se sont largement penchés sur la qualité des voyelles. Une gamme de grandes questions empiriques concernant la variation interne du langage demeurent donc ouvertes. Cependant, il y a eu un tournant récent dans la recherche linguistique vers l'utilisation de grands corpus de multiples communautés vocales rendu possible par les changements d'accès aux données et par les progrès des outils de traitement informatique et de modélisation statistique. Ces améliorations permettent d'aborder de nouvelles questions et de réévaluer les perspectives théoriques.

Cette thèse applique cette approche d'étude à grande échelle à la question de la variation phonétique structurée entre dialectes et locuteurs de l'anglais et du japonais. Plus précisément, la thèse démontre que les dimensions variables sont limitées, indiquant les moyens par lesquels la variation phonétique peut systématiquement être structurée entre les dialectes et les locuteurs.

La première étude examine comment les locuteurs du japonais varient dans la réalisation du voisement des plosives (l'aspiration et le voisement lors de la fermeture) en puisant d'un vaste corpus spontané. Grâce à la modélisation statistique, il a été constaté qu'il existe des tendances fortes dans l'emploi d'indices pour marquer les contrastes malgré l'importante variation individuelle. Ces relations étaient beaucoup plus faibles entre les indices, contrairement à certains travaux antérieurs sur l'anglais et l'allemand, suggérant que la structure de la variation est propre à chaque langue et sensible à la spécification phonologique sous-jacente.

Les deux études suivantes portent sur l'anglais en exploitant un corpus à plusieurs corpus (15 corpus spontanés, 30 dialectes) du projet SPeech Across Dialects of English (SPADE). La deuxième étude de cette thèse explore la variation dialectale et individuelle dans la différence de durée entre les voyelles avant les consonnes. Les résultats ont démontré que l'effet du voisement était bien plus petit dans la parole spontanée que dans la parole de laboratoire. Il a également été observé que la variation entre dialectes est plus large que la variation entre individus d'un même dialecte. Ces résultats suggèrent que l'effet du voisement est à la fois plus subtilement contrôlé et plus variable que précédemment rapporté,

tout en restant remarquablement stable dans chaque communauté.

La troisième étude applique l'approche multi-corpus pour examiner la variation dialectale des voyelles en anglais grâce aux données de 11 corpus (21 dialectes). Cette étude a examiné comment plusieurs propriétés vocaliques (les formants, la trajectoire formantique et la durée) peuvent caractériser la variabilité entre les dialectes de l'anglais. Grâce à aux techniques de classification et de réduction de la dimensionnalité, il a été constaté que toutes les mesures étaient informatives pour localiser la variation vocalique entre dialectes. Le rôle relatif de chaque mesure est propre aux voyelles et aux dialectes, donc la caractérisation optimale diffère selon la voyelle.

Ensemble, ces résultats démontrent l'utilité et le rôle des études « à grande échelle » pour répondre aux questions centrales sur la variation phonétique. En particulier, grâce à l'utilisation de plusieurs corpus et l'application de techniques informatiques et statistiques pour modéliser les tendances dans des données fortement déséquilibrées, il est possible de révéler l'étendue de la variabilité phonétique à plusieurs niveaux de la structure linguistique.

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply thankful to my supervisors Morgan Sonderegger and Jane Stuart-Smith, and my committee members Charles Boberg and Meghan Clayards.

Morgan has long been a supervisor to me, co-supervising my 2nd Evaluation paper prior to this thesis. In this time, he has provided me with his invaluable support and mentorship, and his classes on computational phonology and statistical methods sparked my interest for approaching linguistic analysis from a quantitative perspective. Any skills in statistical analysis I may have are directly as a consequence of Morgan's guidance and encouragement to explore new methods and approaches. I thank him for having faith in me and for giving me the confidence to believe in the value of my research. More than academic supervision, Morgan has always been a source of understanding and kindness, of which I am fortunate to have been a recipient throughout my studentship.

I met Jane at the start my thesis research, and she immediately taught me the value in being excited about the linguistic enterprise. As a supervisor she has taught me much about phonetics and shared invaluable advice on the nature of research, equipping me with the skills to succeed as a researcher. As a host during my research visit, she welcomed me into her laboratory and her home, and inspired me to explore the research ideas that excited me. She is both an exceptional mentor and a valuable friend, and I am grateful for all she has given me.

I would like to thank the McGill University Linguistics Department for everything they have done to support and care for me during my tenure as a student. Deserving of particular thanks are Heather Goad for first taking a chance on me and subsequently supervising my MA research, Francisco Torreira for co-supervising my 2nd Evaluation paper, and Michael Wagner for his academic and professional mentorship, and giving me options and support when I really needed it. Andria De Luca and Giulianna Panetta have done so much for me, helping me navigate complex administrative issues as a result of my often atypical circumstances.

The students of the linguistics department, past and present, have done so much to make my student life enjoyable and full of intellectual stimulation. In particular, I would like to thank Oriana Kilbourn-Ceron, Yeong Woo Park, Jeff Lamontagne, Henrison Hsieh, Hye Young Bang, Yuliya Manyakina, and Justin Royer for their companionship. Additional thanks to Jeff Lamontagne for translating my thesis abstract into French.

Chapters 3 and 4 of this thesis were made possible by virtue of the SPeech

Across Dialects of English (SPADE) project. The success of both the project and this thesis comes from the effort of the SPADE team members and data guardians. Specifically, I would like to thank Michael McAuliffe and Michael Goodale for their work in building and maintaining the software on which SPADE relies and for helping me troubleshoot my many technical problems, Rachel Macdonald and Vanna Willerton for collating and managing SPADE datasets, Jeff Mielke for sharing his expertise on vowel formants, the data guardians for kindly sharing their data analysed in these chapters, and the SPADE research assistants for carefully checking, correcting, and validating datasets.

I was fortunate to spend the autumn of 2019 as a visiting student at the University of Glasgow, supported by funding from the Fonds de Recherche du Québec Société et Culture. This was a fantastic experience, due to the warm welcome I received from both the faculty and students of the department. Special thanks go to Julia Moreno, whose kindness and eccentric humour gave colour to my time in Glasgow. As two students both at the end of our PhD journeys, the absurd and hysterical conversations we have shared have provided both a source of essential support and the bedrock of a lifetime friendship.

I was able to complete this thesis due to the extraordinary friendship of innumerable people. Henry Bushell, as my best friend, has given me companionship when I needed it most, and I will always be grateful to him for allowing me to share my passions. I thank Peter Chen for making my time in Montreal a unforgettable one, for the beer-fuelled discussions of statistics and programming languages, and for always encouraging me to focus on the big picture. I want to thank Emily Simone-Legris for being my cheerleader, for sharing so many extraordinary experiences with me, and for always providing me with care, support, and understanding throughout the writing of this thesis. I also want to thank Alexandre Dorion for being my first and longest friend in Montreal, François Chenard for inspiring me with his passions, Chiao-Ting Fang for giving me much-needed fun and entertainment, and Colin Bath for his contagious sense of curiosity. This thesis would not have been possible without the support and friendship of Julien Havelock-Hill, Lewis Bird, Antonia Sullivan, Matthew Duncan, Sam Henning, Sebastian Scott, Lauren Bishop, Charlie Loveday, Jules Lipton, and Matthew Evans.

I give my constant love and appreciation to my parents and family, whose care, understanding, and support in all of the decisions I have made, and for their generosity and unwavering love that made this thesis possible. This thesis is dedicated to them.

Contribution of authors

The studies presented in this thesis have each been prepared for publication in peer-reviewed journals. I am the primary author of each manuscript.

Chapter 2 has been accepted for publication in the *Journal of the Acoustical Society of America*, and was co-authored with Morgan Sonderegger and Jane Stuart-Smith. I am responsible for conceiving of the study and its research questions, design of the methodology, data collection, and the original draft of the manuscript. Statistical analysis was performed by me in consultation with Morgan Sonderegger. I led interpretation of the results and revision of the manuscript, with input from all authors. A preliminary report of this study was published in the *Proceedings of the 19th International Congress of Phonetic Sciences* (Tanner et al., 2019a).

Chapter 3 has been published in *Frontiers in Artificial Intelligence*, and was co-authored with Morgan Sonderegger, Jane Stuart-Smith, and Josef Fruehwald (Tanner et al., 2020). The research questions were conceived by the first three authors, and the data was sourced as part of the SPeech Across Dialects of English (SPADE) project and Josef Freuhwald. I extracted the data, and performed the statistical analysis with input from Morgan Sonderegger. I wrote and revised the manuscript with input for all authors. A preliminary analysis using a subset of this data was published in the *Toronto Working Papers in Linguistics* (Tanner et al., 2019b).

Chapter 4 has been prepared for submission, and is co-authored with Morgan Sonderegger and Jane Stuart-Smith. The project and its research questions were conceived by me in discussion with the other two authors. The data was sourced as part of the SPADE project, and the data extraction was performed by me. I performed the statistical analysis with input from Morgan Sonderegger. The manuscript was written by me, and all authors have contributed to its revision.

Chapter 1

Introduction

Phonetic variation within languages – differences between dialects and speakers of the same language – has been one of the most substantial areas of research within the phonetic, dialectological, and sociolinguistic literatures. Underlying the study of phonetic variation is the observation that variability is *structured*: the realisation of segments is not determined randomly, but is rather explainable as a function of a set of linguistic, social, and cognitive factors (Liberman et al., 1967; Labov, 1972, 1994, 2001, 2011; Foulkes et al., 2001, 2005; Foulkes, 2010). The implications of structured variability raise a number of interesting questions: for example, considering how multiple sources of variation in the speech signal map to lower dimensions that may aid in the process of speech perception, where speakers need to attend to fewer acoustic-phonetic cues whilst also processing speaker attributes (Liberman et al., 1967; Lisker, 1985; Chodroff and Wilson, 2017; Kleinschmidt, 2018). Similarly, phonetic variation is exploitable, such that speakers can utilise the range of possible realisations to construct and maintain social meaning (Labov, 1963; Eckert, 2012), and construct the ways in which languages undergo sound change (Weinreich et al., 1968; Labov et al., 1972; Labov, 1991; Baker et al., 2011). This thesis addresses these questions by examining the *sources* and *structure* of phonetic variation – in what ways variation may be structured, and by what properties can that structure be explained – and focuses on two kinds of structured variability: variability across *dialects* and across *speakers* of the same language.

The study of variation has largely been carried out through analyses of either single or closely-related speech communities, varying in the use of naturally-occurring and tightly-controlled laboratory speech, and necessarily – given practical constraints on processing and analysing speech – on relatively small numbers of speakers within each community. This thesis presents three studies examining the structure of variability across dialects and across speakers in two languages – English and Japanese – and applies a 'large-scale' approach to the linguistic data and methodology used in each study. For the purposes of this thesis, 'large-scale' refers to the use of data derived from large speech corpora – consisting of hundreds of hours from a number of speakers – and the application of computational techniques to perform automatic measurements and statistical analysis (Vasishth et al., 2018b; Mielke et al., 2019). In this sense, this thesis utilises approaches from both 'corpus phonetics' (Liberman, 2018) and 'corpus sociolinguistics' (Baker, 2010), and follows the increased use of speech corpora in the study of linguistic variation in both phonetic and sociolinguistic research (e.g. Gendrot and Adda-Decker, 2005; Yuan et al., 2007, 2006; Tauberer and Evanini, 2009; Labov et al., 2013; Kendall, 2013; Stuart-Smith et al., 2015; Sonderegger et al., 2017; Tanner et al., 2017).

1.1 Approaches to the study of phonetic variation

This section introduces how the study of structured variability has been approached within both subfields of phonetics (1.1.1) and sociolinguistics (1.1.2), where both fields have focused on different questions concerning the how phonetic variation is structured across dialects and speakers.

1.1.1 Variation in phonetics

Within the phonetic literature, the systematic study of variation arose as a consequence of the ‘lack of invariance problem’ (Liberman et al., 1967; Lisker, 1985): the underlying acoustic-phonetic character of an individual segment, which is presumably essential to the perception of speech, cannot be observed directly, as its realisation is variable as a function of numerous linguistic, acoustic, and cognitive factors. Given the early goals within phonetic research to uncover the fundamental acoustic properties of speech segments, the function of studying variation was to understand how various factors influence the realisation of a segment as a means of reverse-engineering the underlying structure (i.e., the featural specification) of segments and how these structural differences map to acoustic realisation (Liberman et al., 1957, 1967; Raphael, 2005).

Individual differences in phonetic realisation have been recognised in both early anecdotal reports (e.g. Rositzke, 1939; Kenyon, 1940) and experimental studies (Lisker and Abramson, 1964), and the approach to many types of acoustic normalisation procedures attempt to control for individual differences such as gender and age arising from anatomical variation (e.g. Lobanov, 1971; Nearey, 1978; Adank et al., 2004; Clopper, 2009). Differences between speakers have also recently received more direct attention in both studies of speech perception (Magnuson and Nusbaum, 2007; Schultz et al., 2012; Schertz et al., 2015; Kong and Edwards, 2016; Chodroff, 2017; Clayards, 2018a; Kleinschmidt, 2018; Beddor et al., 2018; Kim and Clayards, 2019; Yu and Zellou, 2019) and production (Johnson et al., 1993; Theodore et al., 2009; Clayards, 2018b; Chodroff and Wilson, 2017, 2018; Yu and Zellou, 2019; Sonderegger et al., 2020a). Work has also addressed variability within individual dialects of English (Hillenbrand et al., 1995; Williams and Escudero, 2014), as well as comparing the realisation of vowels across a number of dialects (e.g. Clopper et al., 2005; Fox and Jacewicz, 2009).

1.1.2 Phonetic variation in sociolinguistics

Variation is central to sociolinguistic research, where the analysis of structural differences between dialects (Weinreich, 1954) placed dialectology within the scope of theoretical issues concerning social variation and the systems underlying language change (e.g. Weinreich et al., 1968; Labov et al., 1972; Chambers and Trudgill, 1980). Much work has concerned dialectal variation in vowel systems (Labov et al., 1972; Wells, 1982; Thomas, 2001, 2003; Clopper et al., 2005; Labov et al., 2006) and consonantal alternations (e.g. Labov, 1966; Guy, 1980; Tagliamonte and Temple, 2005). Early sociolinguistic studies analysed and represented the vowel systems of individual speakers as representatives of broader dialectal-patterns (Labov et al., 1972; Labov, 1991), whilst speakers were recognised as having the capacity to vary in their participation in regional patterns in line with numerous factors, including speaker identity, affiliation to community, and the speech context (Labov, 1963). The social role of speaker variability has been further conceptualised in terms of speaker adaption to the social nature of the conversational participants (Bell, 1984), and speakers as agents in the creation of socio-indexical identity and meaning (e.g. Eckert, 2012, 2019).

The dialectological approach within the variationist sociolinguistic tradition formally began with Labov's studies of variation in Martha's Vineyard (Labov, 1963) and New York City (Labov, 1966, 1972), and a substantial number of studies have analysed regional variation in English, focused mainly on Englishes of North America (e.g. Wolfram, 1969; Wells, 1982; Labov, 1991; Clarke et al., 1995; Fridland, 2000; Thomas, 2001, 2006; Labov et al., 2006), the United Kingdom (e.g. Knowles, 1973; Macaulay, 1977; Wells, 1982; Trudgill, 1999; Foulkes and Docherty, 1999; Shackleton, 2007; Kerswill et al., 2008), and, to a smaller extent, on many other dialects of English (e.g. Deterding, 2003; Bradley, 2004; Bekker, 2012; Docherty et al., 2015). The relationship between individuals and their speech community has also been of longstanding theoretical interest to research concerning language variation and change (Guy, 1980; Wolfram and Beckett, 2000; Mendoza-Denton, 2010), where various methodological and scientific issues have been raised by the presence of community-level heterogeneity (Labov, 1966, 1972, 2014; Johnstone, 1996; Schilling-Estes, 2004), including the role of individual variation in the actuation of sound changes (Baker et al., 2011; Stevens and Harrington, 2014; Mielke et al., 2016; Beddor et al., 2018).

1.2 Methods and data in the study of phonetic variation

This section reviews the ways in which studies of dialectal and speaker variation have been carried out, with a focus on the use of instrumental acoustic-phonetic techniques and statistical methods (1.2.1) and the data available to researchers (1.2.2).

1.2.1 Instrumental and statistical methods

Owing to the technical limitations prior to the mid-twentieth century, early experimental research in phonetics utilised relatively small datasets – often a handful of speech tokens – and analysed using impressionistic coding (Jones, 1909). Following the invention of the sound spectrograph (Koenig et al., 1946) and development of the linguistic role of vowel formants (Chiba and Kajiyama, 1941; Fant, 1956), it was then possible to analyse the acoustic properties of speech. During this period, mainly focusing on the pronunciation of words from lists, much work focused on the role of acoustic properties such vowel duration (House, 1961) and formants (Joos, 1948; Peterson and Barney, 1952), and stop voicing (Lisker and Abramson, 1964) in distinguishing individual segments from others within the linguistic system. Studies extending these analyses in controlled connected speech contexts followed thereafter (Umeda, 1975; Crystal and House, 1982), as well as research focusing on suprasegmental variation such as speech rate (Harris and Umeda, 1974; Crystal and House, 1990) and prosodic structure (Browman and Goldstein, 1991; Wightman et al., 1992). More recent research has utilised data from conversational spontaneous speech, often for the purposes of examining processes of phonetic reduction otherwise unobservable in tightly-controlled laboratory settings (e.g. Johnson, 2004; Torreira and Ernestus, 2011; Ernestus and Warner, 2011; Ernestus et al., 2015; Stuart-Smith et al., 2015), along with numerous approaches to statistical modelling of phonetic data (e.g. Williams and Escudero, 2014; Sonderegger et al., 2017).[1]

The use of acoustic-phonetic methods were present in early sociolinguistic research (Labov, 1963, 1966; Labov et al., 1972), in which acoustic analysis was performed on a mixture of both sociolinguistic interview speech and pronunciations from word lists. Following this, the then-emergent field of sociophonetics has applied the methodological and quantitative approach of phonetics to questions regarding socially-grounded language variation and change (Hay and Drager, 2007; Foulkes et al., 2010; Thomas, 2011; Baranowski, 2013). Concretely, this involves the use of instrumental techniques, such as acoustic analyses of vowel formants (e.g,. Clopper et al., 2005; Labov et al., 2006; Jacewicz et al., 2009) and consonantal realisation (e.g. Tagliamonte and Temple, 2005; Foulkes and Docherty, 2006; Schleef, 2013; Temple, 2014; Stuart-Smith et al., 2015), and multifactorial statistical analysis including linear and logistic regression (Schleef, 2013), mixed-effects models (Fruehwald, 2016b), and, more recently, non-linear modelling (Fruehwald, 2013; Cole and Strycharczuk, 2019; Renwick and Stanley, 2020).

[1] See Roettger et al. (2019) for a recent review of previous and current statistical methodology in phonetic research.

1.2.2 Experimental and corpus data

Much of the previous and contemporary research in both fields has predominantly focused on data representing either a single or a small collection of language varieties. This may take the form of comparing similar speech communities which may differ in a single sociolinguistic dimension (e.g. Risdal and Kohn, 2014; Swan, 2016), dialects expected to take part in similar patterns of regional variation (e.g. Farrington et al., 2018), or using data from a specific dialect as a representative example of the language as a whole (e.g. Gay, 1970; Johnson, 2004; Chodroff and Wilson, 2017). Whilst a range of studies have investigated phonetic variation across many dialects (Thomas, 2001; Clopper et al., 2005; Labov et al., 2006; Tauberer and Evanini, 2009), the process of collecting and analysing data from such a variety of regions makes studies exceptional in their dialectal scope, but unfeasible as a methodological norm for phonetic and sociolinguistic research.

The use of speech corpora for phonetic research – such as the TIMIT corpus (Garofolo et al., 1993) – began in the early 1990s (Byrd, 1993; Keating et al., 1994; Sun and Deng, 1995), and has since provided opportunities to reuse linguistic material for multiple distinct research projects, allowing the same data to be re-examined in terms of different linguistic and acoustic phenomena. The focus on language 'in use' within sociolinguistic research has made corpus-based analysis standard since the earliest sociolinguistic and dialectological studies (Labov et al., 1972; Macaulay, 1977), and sociolinguistic speech corpora now exist for a wide range of English dialects (e.g. Bois et al., 2000; Labov and Rosenfelder, 2011b; Labov et al., 2006; Gordon et al., 2007; Stuart-Smith et al., 2017; Kendall and Farrington, 2018). The relatively recent growth and increased access to speech corpora, such as those from the Linguistic Data Consortium (LDC) for English, the National Institute for Japanese Language and Linguistics (NINJAL) and National Institute of Informatics Speech Resources Consortium (NII-

SRC) for Japanese, allow both for validating linguistic theories across multiple heterogeneous datasets and for more easily examining patterns of phonetic variation across a wide spread of dialects. Similarly, the development of speech data management systems (e.g. Rose et al., 2006; Kendall, 2007; Fromont and Hay, 2012; Winkelmann et al., 2017) has made possible the ability to process data from speech corpora with less need for extensive technical skills.

1.3 Integrated corpus analysis

Whilst the advent of speech management systems has greatly decreased the technical overhead involved in performing phonetic analysis with corpus data, the process of repeating the same analysis across *multiple* corpora remains difficult. The SPeech Across Dialects of English (SPADE) project (Sonderegger et al., 2020b, https://spade.glasgow.ac.uk/) focuses on performing 'integrated corpus analysis' – iteratively re-applying the same analytical procedure to multiple corpora – for the purposes of exploring phonetic and phonological variability across dialects of English. This is enabled through the use of the collection of speech data and specifically-designed software for processing, managing, and analysing corpus data: the Integrated Speech Corpus Analysis (ISCAN) software (McAuliffe et al., 2019). Chapters 3 and 4 form part of the SPADE project, where datasets collected via the SPADE project are processed via the ISCAN software, and so the following sections summarise the data collection (1.3.1) and processing (1.3.2) in the context of SPADE, as well as addressing the methodological approaches (1.3.3), challenges, and theoretical caveats associated within SPADE and other large speech corpus projects (1.3.4).

1.3.1 Data collection

The SPADE project, for which the ISCAN software was developed, was devised as a scheme to easily perform large-scale analysis across a range of English varieties. The project relies on the availability of existing speech corpora: no corpus was created with the intent of being included within SPADE; rather, the project utilises a repository of public and private datasets originally created for other purposes. The research goals of SPADE concern phonological variability across time and space, particularly across dialects of North American (Canada, United States), British (England, Scotland, Wales), and Irish English. The corpora collected as part of the SPADE project cover the major dialectal variants of these areas, such as the regional groupings of North America (Thomas, 2001) and Great Britain (Trudgill, 1999), vary across recorded time from the 1960s until the 2010s, and vary in speech style including spontaneous casual speech, broadcast speech, interviews, read speech, and word lists. Some corpora contain speech exclusively from a single dialect: the Sounds of the City corpus (SOTC, Stuart-Smith et al., 2017) and Buckeye corpus (Pitt et al., 2007) for example, are corpora of Glasgow and Central Ohio English respectively. Other corpora, such as Intonational Variation in English (IViE, Grabe, 2004) and the Santa Barbara corpus (Bois et al., 2000) are 'multi-dialectal', containing speakers from a range of different regions. As the SPADE project has access to corpora which also overlap in dialect coverage – both SOTC and the SCOTS corpus (Anderson et al., 2007) contain Glasgow English speech – this opens the possibility of maintaining a 'many-to-many' correspondence between dialects and the corpora that represent them, which would avoid the potential for a given dialect's values to be dependent exclusively on a specific corpus. Whilst this is possible for some dialects with substantial coverage (e.g., Glasgow, East England), the relative sparsity of the data for other dialects means this coverage by multiple corpora is not possible for all dialects analysed.

1.3.2 Data processing

The main goal of the ISCAN software is to allow for the processing and managing of data from speech corpora in such a way that abstracts from the idiosyncratic format of the original dataset. Concretely, this means that data coming from different sources (e.g., aligned with different aligners, created by different users, etc.) can be added and represented in ISCAN without additional processing of the original data (editing annotation files, performing realignment, etc.). ISCAN is built upon functionality from PolyglotDB (McAuliffe et al., 2017b), a Python module developed for storing and processing linguistic data. Data is represented through graph formalism (Bird and Liberman, 1999), and the processing of data consists of four stages. The first, *import*, refers to the process by which raw corpus data (in the form of paired audio and transcription files) are parsed, and the structure of the transcription is mapped into a graph structure. This graph maintains the hierarchical relationships between and within levels of the annotation – individual phonemes belong to a parental 'word' node, and the linear order of phones within a word is maintained. Support is currently available for corpora aligned with a range of forced aligners (Schiel, 1999; Fromont and Hay, 2012; Rosenfelder et al., 2014; McAuliffe et al., 2017a), as well as some corpus-specific idiosyncratic formats (Garofolo et al., 1993; Pitt et al., 2007).

Imported data can be subsequently *enriched* with additional lexical, structural, and acoustic information. For example, phones can be grouped into syllables, lexical information (such as stress patterns and word frequency) or speaker information (such as dialect, age, gender) can be associated with their respective representations, and acoustic measurements (such as vowel formants or speech rate) can be calculated. Once a corpus has been sufficiently enriched, the corpus can then be *queried* at a particular linguistic level. For example, a potential study interested in vowel duration can generate a query concerning the 'phone' level of the corpus. To focus exclusively on vowel phones, filters can be applied.

Filters can restrict the units in the query (e.g., only vowel phones), the values linearly surrounding the unit (e.g., preceded by a consonant), and the position of the unit within the hierarchy (e.g., at the beginning of a syllable, at the end of a word, etc). Queries represent the data in a tabular formant, with a single row corresponding to a single observation (e.g., one vowel token, one formant measurement), and columns associated with that observation (label of the word and phone, speech rate of the observation, name of the speaker, etc). This tabular data can be *exported* as a comma-separated-values (CSV) datafile – common in linguistic research and can be imported into statistical software for further analysis. In order to collate the data from these corpora for the studies using ISCAN (Chapters 3 & 4), each corpus is processed sequentially, and the output CSVs from each corpus is merged into a single 'master' CSV, which is used as input for the statistical analysis.

1.3.3 Data analysis

As the questions addressed in this thesis concern capturing variability at multiple levels (across the population, across dialects, across speakers), the approach to statistical analysis was motivated by the need for models that allow for flexible specifications and are robust to the underlying complexity of the data. Chapters 2 and 3 utilise Bayesian mixed-models: whilst other classes of multifactorial statistical models would also be appropriate for such studies, Bayesian modelling provides a range of advantages for the nature of the analyses reported here. Bayesian modelling differs from other kinds of regression modelling such as mixed-effects models fit with the lme4 (Bates et al., 2015) package (which are currently the standard in linguistic research) in a few respects.[2] As opposed to estimating a single value for a model parameter (e.g., the duration of a vowel), Bayesian models estimate a *distribution* of likely values for that parameter. This

[2] See Vasishth et al. (2018b) for an introduction to Bayesian modelling for phonetic research.

increases the ease with which one can quantify the uncertainty of a particular estimate – similar values can be computed for lme4-style frequentist models (e.g. Fruehwald, 2016a), though these require a distinct interpretation from those derived from Bayesian regression (Vasishth et al., 2018a; Nicenboim and Vasishth, 2016). Given the likely presence of error in the datasets used in this study (1.3.4), being able to easily ascribe uncertainty to a given result is advantageous. Second, Bayesian models return distributions for all model parameters, including both 'fixed' and 'random' effects: this allows for the easy comparison of estimated values at different levels, such as comparing the value for a particular dialect or speaker with the value estimated across all dialects and speakers. Third, Bayesian models provide substantial flexibility in the kind of models that can be fit, including models which estimate values for multiple dependent variables (Bürkner, 2018): this is applied in Chapter 2, where the values of two acoustic cues are estimated within the same model.

1.3.4 Caveats to large-scale speech corpus analysis

In spite of the technical simplicity of merging these datasets (due the uniform output from ISCAN), a number of analytical decisions are necessary in order to appropriately format the data for statistical analysis, which in turn function as caveats on the kinds of empirical conclusions that can be made from the results of each study.

One such decision concerns how to define a dialect for the purposes of these studies. As the data collection is 'opportunistic' and is constrained by the corpora available at the time of analysis, it is not possible to use corpora that are equally matched in size, structure, or diversity of speakers. Moreover, many dialects are represented by multiple speech corpora (for example, both the SCOTS and SOTC corpora contain speakers from Glasgow), and these corpora are amalgamated to constitute a single regional variety. It is not apparent *a priori* that the groups from these corpora represent a homogeneous set of speakers, which in turn limits the kinds of inferences that can be made about the extent of speaker variability in these dialects, and the extent to which dialect-specific values correspond to the patterns observable for that region.

Another consequence of working with data of this kind is that the observations are too numerous to implement any system of manual correction of measures. Each chapter contains at least one measure that was derived though automatic analysis of the speech signal, and some degree of error in these measures is inevitable. The approach taken in this thesis attempts to account for such potential for errors in two ways. First, as the number of observations used in each study is substantially larger than those in traditional studies in phonetics and sociolinguistics, it is possible to apply a relatively strict filtering criteria for observations, discarding data which is unlikely to be accurate. Second, the statistical methodology employed in Chapters 2 and 3 quantify the degree of uncertainty the effects observed from the data (Vasishth et al., 2018a), which, in turn, encourages conservatism in deriving inferences from the results (section 1.3.3).

1.4 Overview

This thesis examines the sources and structure of phonetic variability across dialects and speakers of two languages – English and Japanese – to address the ways in which phonetic variation may be systematically organised across multiple levels of linguistic structure. This is performed through the application of a 'large-scale' methodological approach, using large datasets of natural speech, automatic acoustic measurement, and statistical analysis.

The first study in this thesis (Chapter 2) examines the structure of variability across individual speakers in the realisation of the Japanese stop voicing contrast. Whilst a number of studies on Germanic languages (like English and German) have shown that speakers are highly constrained in their variability in realising stop voicing contrasts in tightly-controlled speech contexts (Chodroff and Wilson, 2017; Bang, 2017; Hullebus et al., 2018), it is less clear how variation may be modulated in spontaneous speech and in a language where the stop voicing contrast utilises a different set of acoustic cues. Examining speaker variability in two cues – Voice Onset Time (VOT) and the degree of closure voicing – in a corpus of spontaneous Japanese speech (Maekawa et al., 2000), it is found that speakers vary in the overall use of each cue (e.g., some have higher average VOT than others), but are strongly correlated in the size of the ratio (i.e., all speakers have similar sized contrasts; speakers with higher voiced VOT also have high voiceless VOT). Weaker relationships are observed *across* acoustic cues, which extends current understanding of structured variability, based mainly on English, and suggest that structured variation across speakers is partially language-specific – particularly in what acoustic cues are employed by speakers to instantiate the phonological contrast.

The second study (Chapter 3) attempts to quantify the variation across dialects and speakers in the English pre-consonantal voicing effect – the duration difference between vowels preceding voiced and voiceless consonants. In spite of the substantial focus on the voicing effect within the phonetic literature (House and Fairbanks, 1953; House, 1961; Klatt, 1976) and the relatively large size of the effect in English compared to other languages (Chen, 1970; Mack, 1982), little is known about how much the size of the effect varies in spontaneous speech, across dialects of English, and across individual speakers of a dialect. Using corpus data constituting 30 English dialects, it was observed that the voicing effect is substantially smaller in spontaneous speech than previously reported for laboratory speech (e.g. House, 1961), and is highly variable between dialects, ranging from near-null values to a near 50% increase in vowel duration. Variation between speakers was shown to be smaller than cross-dialectal variation, suggesting that speakers vary little from their dialect-specific baseline effect size. These results suggest that the English voicing effect is highly variable and sensitive to dialectal and speech context differences, and demonstrates the potential of re-examining previously well-studied variables using new data sources.

The third study (Chapter 4) evaluates how time-dependent properties of vowels contribute to the dimensions in which vowels differ in their realisation across dialects of English. The importance of time-dependent vowel dynamics – the spectral change in the vowel over its timecourse – for distinguishing vowels within a given dialect has been long acknowledged (e.g. Peterson and Barney, 1952) and extensively examined for some dialects (Hillenbrand et al., 1995; Watson and Harrington, 1999). Cross-dialectal studies of vowel formants have either focused on a small number of closely-related varieties (e.g. Fox and Jacewicz, 2009; Farrington et al., 2018) or focused on acoustically static properties of vowels (e.g. Labov et al., 1972, 2006), leaving unclear how dynamic spectral and duration information varies across a wide range of English dialects. Specifically, this chapter considers how acoustic information, such as a vowel's position in formant space, the shape of the formant trajectory, and duration are together required to demonstrate systematic variation in five vowels across 21 English dialects. These measures were evaluated using dialect classification and dimensionality reduction techniques, where it was shown that all measures contribute to the variation of vowels across English dialects. Information regarding formant position was found to play the largest role in distinguishing dialects in both experiments, with trajectory shape and duration providing additional resolution. The role of each measure was also found to be somewhat dependent on the vowel under investigation, suggesting that continuua on which dialects differ varies by the properties of that vowel within the linguistic system.

Chapter 2

Structured speaker variability in Japanese stops: relationships within versus across cues to stop voicing

2.1 Introduction

The acoustic realisation of segments varies substantially across languages, phonological contexts, and speakers. Within a single language, the realisation of a particular segment can differ as a function of phonological context (Cho and Ladefoged, 1999; Cho and McQueen, 2005), speech rate (Allen et al., 2003), and many other linguistic and social factors (e.g. Foulkes et al., 2001). Individual speakers may differ in the realisation of speech sounds because of numerous factors: some speakers are more prone to hyperarticulation of segments (Lindblom, 1990; Johnson et al., 1993), differ in their anatomical characteristics (Peterson and Barney, 1952), or simply arrive at different acoustic targets as a function of probabilistic approximation of the speech sounds in their community (Bybee, 2001; Pierrehumbert, 2001). This kind of speaker-level variability poses a potential challenge for the perception of speech (Kleinschmidt, 2018), where the mapping from values in a multi-dimensional acoustic space to abstract phonological categories (e.g., [+voice], [-high], etc.) is differently realised for individual speakers

(Liberman et al., 1967; Lisker, 1986). How, then, do speakers successfully convey the presence of singular linguistic categories despite individual variation in those categories' realisations? One way in which individual variability may be constrained is by the existence of underlying *structure* in the realisation of speech sounds across speakers: namely, that speakers' individual productions are related in a way that is fundamentally non-random. For example, whilst speakers vary in the realisation of a single acoustic parameter such as Voice Onset Time (VOT) for stops, the differences between individual speakers' VOT values for different places of articulation are highly correlated (Chodroff and Wilson, 2017; Hullebus et al., 2018). Speakers may also show similar kinds of structured variation across *multiple* cues to the production of a speech sound, evidenced by observed covariation in VOT and F0 across voiced and voiceless stops (Bang, 2017; Chodroff and Wilson, 2018; Schultz et al., 2012; Clayards, 2018b).

Beyond a study on Scottish English (Sonderegger et al., 2020a) and a preliminary analysis on American English (Chodroff and Wilson, 2017), most recent research on structured variation across individuals has focused on production in controlled laboratory speech, either isolated words or reading sentences (Chodroff and Wilson, 2017; Hullebus et al., 2018; Schultz et al., 2012; Clayards, 2018b). The phonetic realisation of stop contrasts is known to be 'enhanced' in laboratory speech relative to conversational speech (Lisker and Abramson, 1967; Baran et al., 1977) – for example voiced/voiceless VOT differences are larger – and so it is less clear how variability is structured in less-controlled speech. Examining spontaneous speech alongside more controlled speech may provide new insights into structured speaker variability in phonetic realisation, as for other aspects of speech, such as variability in vowel production (Meunier and Espresser, 2011; Gahl et al., 2012; DiCanio et al., 2015). Our understanding of structured speaker variability is also largely derived from research which has examined languages such as English and German, which primarily use VOT to signal a range of contrasts in word-initial stops (e.g. Lisker and Abramson, 1964, 1967). How speakers vary in languages where the stop contrasts involve the use of additional phonetic cues is not well understood.

This study addresses these theoretical gaps by focusing on the acoustic realisation of stops in spontaneous Japanese. Japanese uses both *positive* VOT – the period encompassing the duration of aspiration and the stop burst – and the presence of voicing in the stop closure for marking the contrast between voiced and voiceless stops (Shimizu, 1996; Tsujimura, 2014, Section 2.2.1). Typically 'VOT', in work on Japanese and other languages, is defined as the time between the release of the stop and onset of glottal pulsing for the following vowel: VOT is positive if voicing begins after the release of the stop closure, and negative otherwise. In that definition, VOT is both an indirect measure of 'burst duration' and aspiration (when positive) and the presence of voicing during the closure (when negative). In this study, which focuses on structured variability, it is important for us to capture the complex interplay between laryngeal and supralaryngeal actions/timing in Japanese stops through two dimensions. In line with several recent studies which distinguish between positive VOT and the presence of voicing during closure (Kim et al., 2018; Kleber, 2018; Seyfarth and Garellek, 2018; Sonderegger et al., 2020a), we use the term 'pVOT' to refer to the duration of 'burst plus aspiration' following the release of the closure. We use 'voicing during closure' (VDC) to refer to any voicing throughout the stop closure. The Japanese stop voicing contrast has been observed to be undergoing change through the decreased use of voicing during closure, resulting in a system resembling an English-style aspiration contrast (Takada, 2011; Takada et al., 2015), and so may provide some insight into how speakers vary in the use of both pVOT and the degree of voicing during stop closure, as well as in how both parameters are used to realise the voicing contrast. This study expands the search for structured speaker variability by examining the evidence for three kinds of such structure across speakers of spontaneous Japanese: (1) *within* a phonetic cue across different voicing categories (e.g., pVOT between voiced and voiceless stops); (2) the size of the voicing contrast *across* cues (i.e., the relative difference in voiced and voiceless stops); and (3) *across* phonetic cues across and within voicing categories (i.e., the relationship between pVOT and voicing during closure in voiced and voiceless stops).

2.2 Background

2.2.1 Acoustic cues to stops & stop voicing

VOT as traditionally defined, is well-established as the primary acoustic cue for the stop voicing contrast in a range of languages where voiced stops have shorter average VOT than their voiceless counterparts (Liberman et al., 1958; Lisker and Abramson, 1964; Abramson and Whalen, 2017). Japanese maintains a two-way stop voicing contrast, distinguishing between 'voiced' {/b/, /d/, /g/} and 'voiceless' {/p/, /t/, /k/} categories: acoustically, Japanese voiced stops may be realised either with prevoicing (negative) or short-lag VOT (Shimizu, 1996; Nasukawa, 2005; Gao and Arai, 2019), and voiceless stops are realised with a VOT intermediate between short ('unaspirated', Tsujimura, 2014) and long-lag ('moderately aspirated', Shimizu, 1996; Riney et al., 2007). Whilst less is known about variability in Japanese stops, much work has focused on how stops are modulated across languages; assumed that these factors are to some extent language-independent and are thus also relevant for Japanese stops. Stop VOTs are affected by a range of linguistic factors, such as place of articulation (Lisker and Abramson, 1964; Docherty, 1992), preceding phoneme manner (Docherty, 1992; Yao, 2009), vowel height (Klatt, 1975), phrasal position (Lisker and Abramson, 1964; Cho and Ladefoged, 1999; Yao, 2009; Kim et al., 2018), and speech rate (Allen et al., 2003). Most work on English VOT has used controlled speech, though the few studies which have looked at variation in English spontanous speech have confirmed a robust difference in VOT between voiced and voiceless stops (Baran et al., 1977; Yao, 2009; Sonderegger et al., 2017; Stuart-Smith et al., 2015; Sonderegger et al., 2020a).

The degree of vocal fold vibration during the closure (Lisker, 1986), reflected in our VDC measure, is much less studied than VOT, though English voiced stops are more likely to contain VDC than their voiceless counterparts (Docherty, 1992; Sonderegger et al., 2020a). Most research on VDC has focused on English read speech (e.g. Davidson, 2016, 2018; Kim et al., 2018). For both voiced and voiceless stops, VDC is more likely in phrase- or word-medial contexts (Docherty, 1992; Lisker and Abramson, 1964, 1967). VDC in phrase-initial stops, sometimes referred to as 'negative VOT', has been observed for English (Lisker and Abramson, 1964, 1967; Hunnicutt and Morris, 2016) and other languages (Abramson and Whalen, 2017). Additionally, VDC is more likely when the preceding segment is voiced (Docherty, 1992; Davidson, 2016, 2018), also in spontaneous speech (Sonderegger et al., 2020a). With the exception of geminated consonants, all syllables in Japanese are either open (ending in a vowel) or have a nasal coda (Tsujimura, 2014): all segments preceding stops in these cases are underlyingly voiced, then, and this should affect the likelihood of a stop being realised with VDC. Closure voicing is also used as a contrastive cue for voicing in Japanese, though recent studies have shown that the prevoiced variant of the voiced stop has become less common in phrase-initial position (Gao and Arai, 2019), and may represent a sound change towards the exclusive use of positive VOT coupled with F0 variation to signal the voicing contrast (Takada, 2011; Kong et al., 2014; Takada et al., 2015; Gao et al., 2019; Gao and Arai, 2019).

2.2.2 Individual speaker variability in stops

Differences between individual speakers have been noted since the earliest acoustic studies of stop production (e.g. Lisker and Abramson, 1964). As opposed to being random variation, these differences between speakers are highly structured: speaker differences in VOT are consistent after controlling for other linguistic factors, such as speech rate (Allen et al., 2003; Theodore et al., 2009). Speaker mean VOTs for different places of articulation in voiceless stops have been shown to be highly correlated in both English (Chodroff and Wilson, 2017) and German (Hullebus et al., 2018): despite overall differences in a given speaker's mean VOT, realisation of the contrasts between voiceless stops (i.e., /p/ ∼ /t/, /p/ ∼ /k/, /t/ ∼ /k/) exhibits strong linear relationships. With respect to speaker variability across *multiple cues* to stop production, Chodroff and Wilson (2018) show that American English speakers covary in use of three cues (VOT, F0, and spectral centre of gravity), and Glaswegian English speakers covary in the relationship between positive VOT and the degree of VDC (Sonderegger et al., 2020a). Similar relationships exist between VOT and F0 in marking the laryngeal contrast in English, German, and Korean (Schultz et al., 2012; Bang, 2017), whilst Schertz et al. (2015) observed speaker differences in the correlated use of VOT, F0, and closure duration in L2 English-Korean speakers, and Clayards (2018b) reported similar findings for VOT, F0, and following vowel duration in English.

In order to characterise the sources of structured variability within an individual's phonological grammar, Chodroff and Wilson (2017, 2018) propose a 'principle of uniformity'. Here, uniformity refers to a linear relationship in the acoustic production of two segments across speakers; the degree of variation in the difference between two speech sounds across speakers is constrained such that the realisation of one sound has a predictive relationship with the other. Whilst speakers may vary in their overall use of a given phonetic cue (i.e., where that speaker is situated on this line), the relative difference between two segments with respect to that parameter is consistent across speakers. Much of the evidence for Chodroff & Wilson's proposition of uniformity is derived from studies on controlled forms of English, which uses an aspiration-based phonetic implementation of stops.

By examining the structure of speaker variability in spontaneous Japanese, a new language with a different phonetic implementation of voicing, we can consider further possible evidence for phonetic uniformity in a new empirical setting. This examination takes two forms here: the first considers how speakers modulate the stop voicing contrast within a given phonetic cue (pVOT or Voicing During Closure). The second concerns how these two cues are manipulated together in signalling this contrast. Whilst some research has examined speaker variability across multiple cues, especially in English (e.g. Clayards, 2018b; Chodroff and Wilson, 2018), the predictions are less clear for a language like Japanese where the cues to stop voicing differ from English and where a number of possibilities exist. For example, if pVOT and Voicing During Closure share an intrinsic articulatory link, we could expect strong correlations between pVOT and Voicing During Closure, such that speakers with more aspirated stops also produce less Voicing During Closure. This would correspond to the intuition behind the traditional 'VOT' measure, that stop production is often well-characterised by a single dimension (Abramson and Whalen, 2017, a closure voicing–degree of aspiration continuum). Alternatively, the lack of an intrinsic link between the cues may result in no observed correlations between the respective use of pVOT and Voicing During Closure. These questions also address the extent to which phonetic uniformity across speakers might be constrained and whether such constraints may relate to language-specific properties.

2.3 Methods

2.3.1 Data

The data used here comes from the Core subset of the Corpus of Spontaneous Japanese (CSJ, Maekawa et al., 2000), constituting approximately 45 hours of speech recorded 1999-2001 from 137 speakers (58 female), born between 1930 and 1979. Within the CSJ, speaker birth years are grouped into increments of 5 years (e.g., 1930-34, 1935-39, 1940-44, etc); in order to ensure sufficient numbers of speakers per group, speakers were allocated into groups of 10 years (1930-39, 1940-49, etc). The variety of Japanese in the CSJ is 'Common' Japanese: a standard variety that derives many of its linguistic features from the Tokyo dialect (Maekawa et al., 2000). Each recording is approximately 30 minutes long, and is predominantly academic interviews and informal public speaking, though a subset (approximately 5%) is conversational dialogue and reading passages. The Core subset contains extensive phonetic and prosodic annotation, including hand-corrected segmental boundaries, presence of vowel devoicing, and voice quality (Kikuchi and Maekawa, 2003). Relevant for the measures taken here, stops were annotated for (1) onset of stop closure, (2) stop burst – the first transient spike – and (3) the onset of the vowel. The segmentation criteria for the hand correction are provided in Fujimoto et al. (2006): for our purposes, onset of following vowel was determined by CSJ annotators as the beginning of periodicity for the vowel (Fujimoto et al., 2006, p.330); see Figure 2.1. The annotations also noted whether the stop was fully realised, defined by whether a clear closure, burst, and voice onset could be visually observed (the CSJ does not contain annotation for negative VOT).

In order to ensure that stops examined in this study were fully realised, certain stops were excluded from further analysis: any stop marked as not having a clear closure and burst (56,661 tokens); stops followed by a devoiced vowel, as voicing onset could not be ascertained (11,939 tokens); stops immediately following hesitations (11,991 tokens); geminate stops (19,785 tokens), as geminates in Japanese are not phonologically contrastive for voicing in native words and often devoice (Kawahara, 2015); stops from word-medial contexts (72,681 tokens), as stops reduce in these contexts (Cho and Ladefoged, 1999; Kim et al., 2018); and stops from non-spontaneous speech (4,790 tokens). Prosodic position is defined in the corpus using the X-JToBI prosodic-labelling scheme (Maekawa et al., 2002), which numerically represents the perceived strength of a prosodic juncture through 'Break Indices' (BIs). BI labelling is based on a range of perceptual cues including segmental lengthening, F0 reset, and changes in voice quality (Venditti, 2005). Junctures with a BI value of 1 typically represent a word boundary within an Accentual Phrase (AP), a BI value of 2 represents the boundary between two APs, whilst BI values of 3 indicate the edge of an Intonational Phrase (IP). We excluded all tokens with *no* BI value (which are predominantly word-medial). The final set of stops analysed therefore constitutes word-initial stops excluding potentially-problematic cases.

2.3.2 Voicing during closure (VDC)

The goal of the VDC measure is to characterise the presence of voicing during closure, which plays a key part in signalling phonological voicing in Japanese. It is well known, however, that realisation of voicing within the stop closure is more complicated in connected speech than that in isolated words (Lisker and Abramson, 1964, 1967; Abramson and Whalen, 2017). Voicing may continue for the entire stop closure ('full voicing'), or may subside ('bleed') and/or return just prior to the release ('trough') (Davidson, 2016). Cases like this make the traditional definition of 'negative VOT' difficult for characterising the voicing pattern. Davidson (2016, 2018) observed that voicing during closure corresponding to negative VOT in American English appeared in only a handful of to-

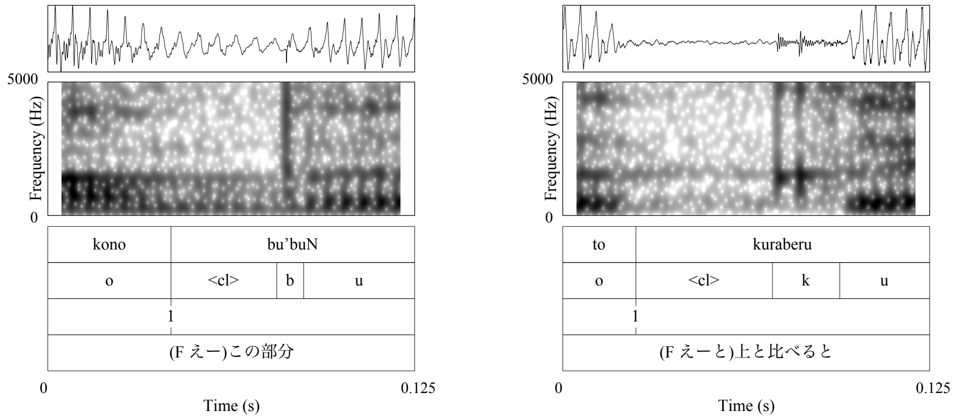

Figure 2.1: Waveforms and accompanying annotations for phrase-internal stops realised with and without voicing during closure ('kon**o** **b**ubun', (left); 't**o** **k**uraberu', (right), respectively) produced by a female speaker taken from a 125ms time window. Closure annotated as <cl>. Top tier represents word-level transcription, second tier contains phone & sub-phone annotations, third tier marks prosodic boundaries via Break Index, and fourth tier contains utterance transcription.

kens. Whilst several studies have focused on negative VOT in laboratory speech (Takada, 2011; Takada et al., 2015; Gao et al., 2019; Gao and Arai, 2019; Kong et al., 2014), no work to our knowledge has examined variability in stop closure voicing patterns in Japanese connected speech similar to Davidson (2016, 2018) for English.

Davidson (2016) notes the likelihood of voicing during closure in English is closely tied to the voicing of the preceding segment: preceding voiced segments (vowels, sonorants) are more likely to induce voicing during closure than voiceless segments. This is important here since *all* preceding segments are voiced: Japanese syllables are either open (i.e., consonant-vowel) or contain a nasal coda (Tsujimura, 2014): as geminated stops have been excluded, all stops are preceded by a vowel or a nasal (potentially with an intervening pause). A preceding vowel does not guarantee the realisation of voicing in the stop closure, however: Figure 2.1 (left) shows a voiced stop realised with voicing throughout the whole stop closure ('full voicing'), whilst no such voicing during closure is evident in a voiceless stop in the same phonetic context (Figure 2.1, right).

Our goal for the VDC measurement is to characterise the presence of phonetic voicing during closure in terms of the likely presence of an active voicing gesture (Beckman et al., 2013). In order to capture this, the presence of VDC is defined in binary terms between the *presence* or *absence* of active voicing during closure. This aims to exclude common cases of passive voicing which are often short (less than 20ms) and weak in amplitude, in contrast to an active voicing gesture, characterised by clear periodic voicing for a substantial portion of the closure and the presence of pitch. This deviates from previous studies on English using similar approaches (Davidson, 2016; Sonderegger et al., 2020a) where voicing during closure was trichotomised into 'no', 'partial', or 'full' voicing, determined by the relative portion of the observed voicing within the closure. The decision to use a binary voicing distinction in this study was based on the goal of restricting to cases where an active voicing target was clearly present or not, as well as on the empirical observation that both Davidson (2016) and Sonderegger et al. (2020a) found that effects were more apparent in their respective binary ('no' versus 'full') models than comparing relative degrees of voicing. Our characterisation of VDC as distinct from pVOT enables both voicing presence and pVOT to be examined as independent cues to stop production: given observations that it is possible for speakers to produce stops with both voicing during closure and pVOT (Abramson and Whalen, 2017; Kim et al., 2018; Sonderegger et al., 2020a), it is important to know if speakers are able to modulate both pVOT and voicing during closure independently to signal the Japanese stop voicing contrast.

In order to calculate a measure of VDC, both the mean F0 and the 'fraction of unvoiced frames' were extracted from the labelled stop closure using Praat Voice Report (Boersma and Weenink, 2017). As Voice Report has been known to produce inaccurate measurements in specific circumstances, our calculations fol-

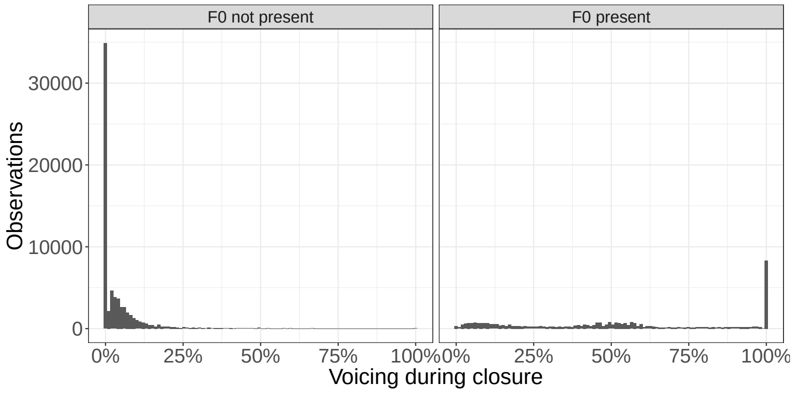

Figure 2.2: Histograms showing the distribution of the percentage of voicing during closure by whether F0 was also detected within the stop closure. 100 bins used within each histogram, meaning that each bar represents 1%.

lowed Eager (2015): specifically, the Voice Report was produced by a Praat script without using the Editor window, using gender-specific pitch ranges (70-250Hz for males; 100-300Hz for females), and a time step of 0.001 seconds. The percentage of voicing during closure was calculated by subtracting 100 from Voice Report's proportion of the interval with *no* voicing: for example, if Voice Report returned an unvoiced closure value of 66%, then voicing % = $100 - 66 = 34$.

Our main goal involved determining which instances of stop voicing were most likely produced with an active voicing gesture. For the purposes of this study, tokens which satisfied two criteria were analysed. The first was whether F0 was present in the closure; the second was whether a significant portion of the closure contained voicing. Numerous values have been proposed in the literature for what proportion of the closure reflects active voicing, such as 'greater than 50%' (Abramson and Whalen, 2017) and 'greater than 10%' (Davidson, 2016). Here, decisions regarding the cutoffs were determined by examining the distribution of voicing during closure percentages with and without the presence of F0. As shown in Figure 2.2, voicing during closure with no accompanying F0 (left panel) ranges from 0% to approximately 15%, and so VDC (reflecting an active voicing gesture) was considered to be absent for such tokens. When F0 is present (right panel), a large number of tokens exhibited 100% voicing during closure with a small cluster around 50%. To include these tokens, the 'present' VDC category was defined as tokens with the presence of F0 and at least 35% voicing in the closure. Other cases were taken to indicate that voicing was unreliable: F0 may have been present but the lack of substantial voicing % suggests potential voicing bleed. Unreliable tokens were excluded (18,960; 17.5%), meaning that all remaining tokens are assumed to be realised with either no voicing during closure or an active voicing gesture. Our final dataset used for analysis contained 90,160 tokens (3,440 types) from 137 speakers (58 female), corresponding to an average of 658 tokens per speaker (range of tokens per speaker: 149–2,913).

2.3.3 Model

The goal of this study is to examine evidence for structured speaker variability (1) within individual acoustic cues; (2) in the voicing contrast across cues; and (3) across cues within individual phonetic categories. In order to address these questions, pVOT and VDC were statistically modelled to characterise individual speaker differences whilst controlling for a range of factors known to influence both cues (Section 2.2.1). pVOT (log-transformed) and VDC were jointly modelled using a multivariate Bayesian mixed model fit with *brms* (Bürkner, 2018), an R front-end for the Stan programming language (Carpenter et al., 2017). A Bayesian model returns a *distribution* of potential values for all model parameters, which makes it possible to estimate correlations across speakers as well as the uncertainty associated with each correlation. This is ideal for addressing all three research questions, as it means that the strength of relationships across speakers can be characterised formally in terms of both the strength of the correlations and the range of possible correlations consistent with the data. As both pVOT and VDC are fit within the same model, it is possible to also directly estimate the speaker correlations *across* phonetic cues, which is crucial for research questions (2) and (3). Finally, the use of a statistical model to estimate speaker correlations, rather than estimating correlations from empirical data as in most previous work on structured speaker variability, allows for correlations (and individual speaker values for each cue) to be estimated whilst controlling for the range of other factors known to affect both pVOT and VDC (Sec. 2.2.1).

The model consists of a sub-model predicting pVOT and a sub-model predicting VDC, and terms linking these sub-models together. We first describe the terms in each sub-model, which were identical. Each sub-model included the following population-level ('fixed-effect') predictors for stop **voicing**, previous phoneme **manner**, speaker **birth year** and **gender**, stop **place of articulation**, speech **style**, prosodic **position**, log-transformed word **frequency**, speaker **mean** and **local** (relative to mean) speech **rate** (Sonderegger et al., 2017; Stuart-Smith et al., 2015), the presence of a preceding **pause**, and following vowel **height**. To control how each predictor influenced the realisation of the voicing contrast, two-way interaction terms between stop voicing and all other predictors were also included in the model. Continuous predictors (speaking rates, frequency, vowel duration) were centred and divided by two standard deviations (Gelman and Hill, 2007). Two-level factors (voicing, accent, gender, vowel height, pause) were converted into binary (0/1) measures and centred. Predictors with three or more levels (birth year, place of articulation, phoneme manner) were coded with sum contrasts. For group-level ('random-effect') predictors, the model was fit with a random intercept for words; speaker-level effects consisted of a random intercept and random slopes for all population-level predictors (with the exception of style, age, and gender). As the relationship between a speaker's overall value for pVOT/VDC and the size of their voicing contrast is of direct interest, both models included a correlation term between the speaker-level intercept and the voicing predictor. The pVOT and VDC sub-models were tied together by three correlations between the key speaker-level effects: intercepts, voicing, and the correlation between them. For example, the correlation term between the pVOT intercept and the VDC intercept captures the extent to which speakers with higher mean pVOT are more likely to use VDC. The model used 8000 samples across 4 Markov chains and was fit with weakly-informative 'regularising' priors (Nicenboim and Vasishth, 2016; Vasishth et al., 2018b) of normal distributions with a mean of 0 and standard deviations of 1 and 0.5, and 0.5 for pVOT intercept, VDC intercept, and fixed effect parameters respectively. The default prior in *brms* for group-level effects was used: a half Student's *t*-distribution with 3 degrees of freedom and a scale parameter of 10. Correlations used the LKJ prior (Lewandowski et al., 2009) with $\zeta = 2$, in order to give lower prior probability to perfect (1/-1) correlations, as recommended by Vasishth et al. (2018b).[1] All data and code used is available at `https://osf.io/grw25/`.

2.4 Results

The research questions concern the relationships observed across speakers *within* each cue (1) as well as *across* both cues (2, 3), and so correlations were calculated for each of the 8000 draws from the posterior sample and reported as the median, 95% credible interval (CrI), and the posterior probability of the parameter not including 0, using `fitted_draws` and `median_qi`, respectively, from the *tidybayes* package (Kay, 2019). Speaker-level variability is first examined *within* pVOT and VDC separately (2.4.1) before examining the relationships *between* both cues across speakers (2.4.2). Following the suggestions of Nicenboim and Vasishth (2016), we consider there to be strong evidence for a non-null effect if the 95% CrI for the parameter does not include 0; if 0 is within the 95% CrI but the probability of the parameter not changing direction is at least 95%, this is considered

[1] To ensure that the correlations reported were not due to the choice of a specific prior, an identical model with a weaker 'flat' prior ($\zeta = 1$) was also fit. The correlations estimated from this model, of primary interest for our research questions, were near identical (within 0.01) to those from the stronger model, indicating that the evidence for the correlations in the data is strong enough not to be affected by the subjective choice to use a more informative prior.

Table 2.1: Median correlation, 95% credible intervals (CrI), and posterior probability of within-cue correlations (Spearman's ρ) across speakers sampled from the model posterior with all other predictors held at their 'average values' (e.g., mean word frequency, mean across all places of articulation, etc).

Correlation	ρ	95% CrI	$\Pr(\rho <> 0)$
Voiceless pVOT, Voiced pVOT	0.77	[0.709, 0.821]	1
Voiceless VDC, Voiced VDC	0.664	[0.594, 0.729]	1

to represent weak evidence for a given effect. Crucially the strength of evidence for an effect is distinct from its magnitude, and so the strength of a given predictor's effect on pVOT/VDC is considered alongside its relative evidence. The size or magnitude of a given correlation is assessed in terms of Cohen's conventions (Cohen, 1988): correlations with sizes between 0 and 0.1 (in either direction) are considered to be *negligible*; those with sizes between 0.1 and 0.3 to be *small*; between 0.3 and 0.5 to be *medium*; and *strong* correlations have values larger than 0.5. Cohen's conventions are considered to be heuristic and should be considered relative to previous effect sizes observed for a given phenomenon. Given the relative scarcity of results on the relationships across speakers, Cohen's conventions provide some initial benchmarks against which to evaluate the relative relationships within and across phonetic cues.

2.4.1 Within-cue variability

The effects of the population-level parameters on pVOT were as expected, including the size of the voicing contrast (Table 2.3 in Section 2.7). As the pVOT voicing contrast is maintained across all population-level effects (i.e., no parameter neutralised or reversed the basic voiceless > voiced pattern, including speaker age) and speaker-level variability is of primary interest for our research questions, these parameters provide controls for the speaker-level variability; the fixed effects are not discussed further. Figure 2.3 (left) demonstrates the strong correlation between speakers' voiced and voiceless pVOTs (95% CrI = [0.709,

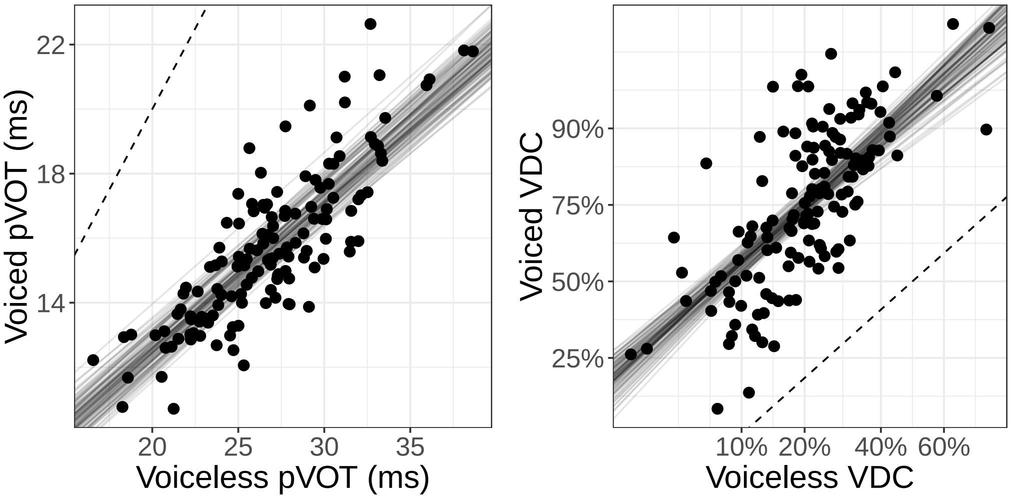

Figure 2.3: Model-estimated cue values for pVOT (left) and VDC (right) for voiceless (x-axis) and voiced (y-axis) stops. One point is the posterior mean value for a particular speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation. Dashed line is $y = x$, where the value for voiceless stops equals that for voiced stops. pVOT plot in linear (millisecond) scale; VDC plot is in logit-scaled probability scale to illustrate differences at extreme upper and lower probabilities.

0.821]; Table 2.1, row 1): each point represents a speaker's median estimated voiceless (x-axis) and voiced (y-axis) pVOT value. All individual speakers have higher pVOTs for voiceless than voiced stops, indicated by all points appearing on one side of the dashed $y = x$ line. Speakers differ in their particular pVOT values, but the relative difference between their voiced and voiceless pVOTs (i.e., the voicing contrast) is consistent: the regression lines demonstrate this linear relationship, where speakers both maintain the contrast between stops, and speakers with long pVOTs for voiceless stops also have long pVOTs for voiced stops.

As for VDC, no population-level effect neutralised or reversed the VDC voicing contrast (Table 2.4 in Section 2.7), meaning that VDC is always predicted to be more likely for voiced than voiceless stops ($\hat{\beta}$ = 2.99, CrI = [2.76, 3.21], Pr($\hat{\beta}$ $>$ 0) = 1). Note, however, the large effect of the presence of a preceding pause on VDC, which suggests that speakers producing spontaneous Japanese are substantially less likely to produce VDC directly following a pause ($\hat{\beta}$ = -3.24, CrI

Table 2.2: Median correlation, 95% credible intervals (CrI), and posterior probability of across-cue correlations (Spearman's ρ) across speakers sampled from the model posterior with all other predictors held at their 'average values' (e.g., mean word frequency, mean across all places of articulation, etc). pVOT contrast = voiceless pVOT − voiced pVOT; VDC contrast = voiced VDC − voiceless VDC.

	Correlation	ρ	95% CrI	$\Pr(\rho <> 0)$
Voicing contrast	pVOT contrast, VDC contrast	0.198	[−0.001, 0.346]	0.974
Within-category	Voiced pVOT, Voiced VDC	−0.348	[−0.423, −0.27]	1
	Voiceless pVOT, Voiceless VDC	0.135	[0.038, 0.228]	1
Across-category	Voiceless pVOT, Voiced VDC	−0.152	[−0.233, −0.066]	0.99
	Voiced pVOT, Voiceless VDC	0	[−0.092, 0.093]	0.5

= [−3.51, −2.97], $\Pr(\hat{\beta} < 0) = 1$), consistent with experimental findings (Gao and Arai, 2019). Comparing across voicing categories, Figure 2.3 (right) illustrates that speakers maintain a strong positive relationship between their voiced and voiceless VDCs (95% CrI = [0.594, 0.729]; Table 2.1, row 2). No speaker has a reversed voicing contrast for VDC, reflected by all speaker values (represented as points) appearing above the $y = x$ line. The consistent positive slope of the regression lines illustrate that, as with pVOT, speakers who are more likely to produce VDC for voiced stops are also more likely, on average, to produce voiceless stops with VDC.

2.4.2 Across-cue variability

Having shown above how speakers vary *within* a single cue (pVOT, VDC) between voiced and voiceless stops (question 1) we now address whether speakers vary *across* cues in production, where speakers may coordinate both cues in signalling the stop voicing contrast (question 2), or specific segments (question 3). Comparing the size of the voicing contrast for each cue, a weak positive relationship across speakers can be observed (95% CrI = [−0.001, 0.346]; Table 2.2, row 1): this can be interpreted as meaning that the voicing contrast sizes across cues are somewhat linked, with speakers differing in precisely how they realise the voicing contrast simultaneously across both pVOT and VDC (Figure 2.4).

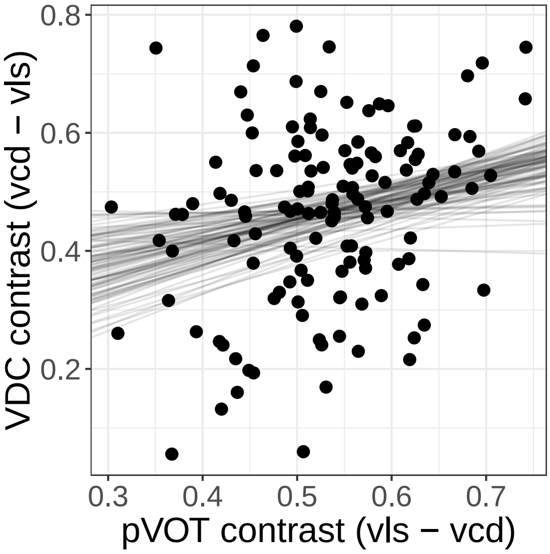

Figure 2.4: Model-estimated voicing contrast sizes for pVOT (x-axis) and VDC (y-axis). Each point is the posterior mean for a particular speaker. Black lines are 100 lines of best fit drawn from the model posterior to show direction and uncertainty in the correlation.

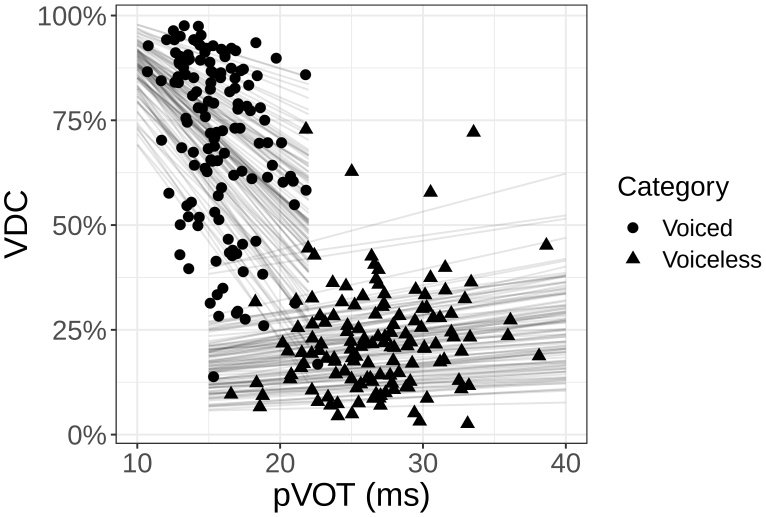

Figure 2.5: Model-estimated cue values for pVOT (x-axis) and VDC (y-axis). Voicing category of the stop is represented by shape (points = voiced; triangles = voiceless). Points and lines represent the same values as in Figures 2.3 and 2.4.

Given the strong correlations across speakers in single use of a given cue (Figure 2.3) and the observation that speakers only weakly vary in the size of their voicing contrast across both cues (Figure 2.4), the question remains as to how speakers covary in the use of pVOT and VDC within specific phonetic categories. In other words, do speakers' values for one cue (e.g., pVOT) within a category (e.g., voiceless stops) correlate with their values for the other cue (VDC) in that same category? Figure 2.5 demonstrates this combination of cues by voicing categories, and illustrates an asymmetry in the pVOT-VDC relationship between voiced and voiceless stops. Speakers provide strong evidence for a negative relationship of medium strength between pVOT and VDC in voiced stops (Figure 2.6, top-left), meaning that speakers with larger voiced pVOTs have a lower voiced VDC likelihood (95% CrI = $[-0.423, -0.27]$; Table 2.2, row 2). For voiceless stops, however, there is strong evidence for a weak *positive* relationship (95% CrI = [0.038, 0.228]; Figure 2.6, bottom-left; Table 2.2, row 3). A negative relationship is also observed between speakers' voiced VDC rate and their voiceless pVOTs, though this is much smaller in magnitude than the voiced pVOT-voiced VDC relationship (95% CrI = $[-0.233, -0.066]$; Figure 2.6, bottom-right; Table 2.2, row 4); voiceless VDC does not show a meaningful correlation with voiced pVOT across speakers (95% CrI = $[-0.092, 0.093]$; Figure 2.6, top-right; Table 2.2, row 5).

2.5 Discussion

The phonetic realisation of segments differs across languages, dialects, phonetic contexts, and individual speakers. Recent research has observed that this variability across individual speakers is *structured*: whilst speakers may differ in the overall value of a particular phonetic cue, they may demonstrate covariation in the use of one or more cues to mark linguistic contrasts (e.g. Theodore et al., 2009;

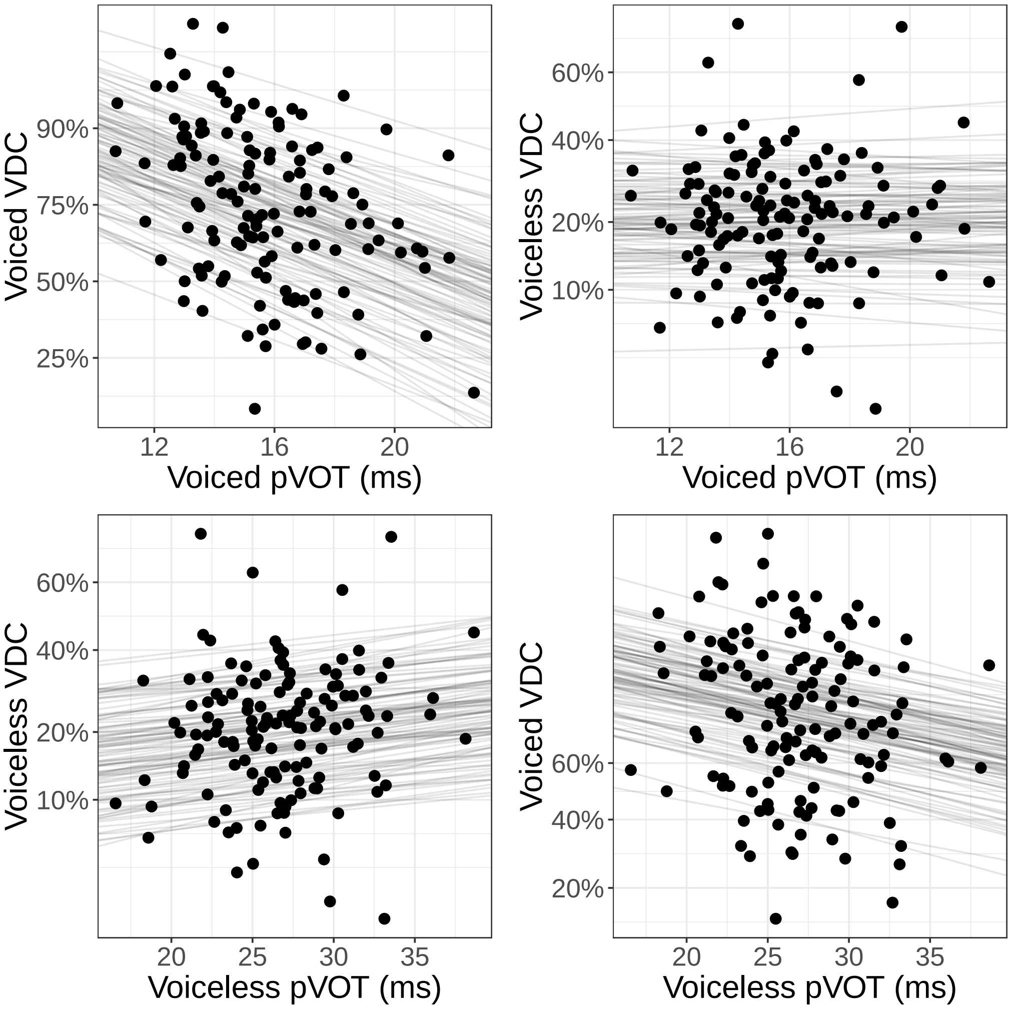

Figure 2.6: Model-estimated cue values for pVOT (x-axis) and VDC (y-axis), comparing relationship between cues either within (left) or across (right) a given stop category. Points and lines represent the same values as in Figures 2.3 and 2.4. pVOT in linear (ms) scale; VDC in logit-scaled probabilities to show differences at extreme probabilities (near 0% or 100%).

Chodroff and Wilson, 2018; Sonderegger et al., 2020a). Much previous empirical work on structured speaker variability has focused on controlled English and German speech: it is not known how speaker variability may be structured in a language that shows different phonetic and phonological signalling of linguistic contrasts. This study begins to address these empirical gaps by examining positive VOT and VDC as cues to word-initial stop voicing in spontaneous Japanese. Strong within-cue relationships are observed across speakers between voiced and voiceless stops: whilst speakers differ in their overall values of pVOT or VDC, speakers are consistent in the relative *difference* between pVOT or VDC in marking the voicing contrast. These within-cue relationships are of comparable magnitude to the strongest correlations observed for English stops in both laboratory (Chodroff and Wilson, 2017, 2018) and spontaneous English (Sonderegger et al., 2020a), demonstrating that structured speaker variability is present in laryngeal systems beyond English aspiration-type systems, and in more than one independent cue to a contrast in spontaneous speech.

Here, most of the predictable variability across individual speakers is *within* a given phonetic cue (2.4.1), as compared with variability *across* the two cues (2.4.2): no across-cue relationship (Table 2.2) is as strong as either of the within-cue correlations (Table 2.1). The size of the voicing contrasts between pVOT and VDC is positively correlated across speakers (Figure 2.4). This could be evidence that speakers vary in the degree of 'clarity' in their speech: speakers align multiple cues to a voicing contrast simultaneously in order to maximise the acoustic distinctiveness between the categories, as opposed to emphasising one cue over another (Bang, 2017; Clayards, 2018b). An explanation in terms of speech clarity does not straightforwardly apply in this data, however, for two reasons. First, the size of the correlation itself is small (Table 2.2, row 1), reflecting only a weak relationship between the two cue contrast sizes. Second, this predictive pattern for the use of pVOT and VDC is observed only for voiced stops: whilst the pVOT-VDC relationship is negatively correlated in voiced stops, no clear relationship is observed for voiceless stops (Table 2.2; Figure 2.5). This suggests that the pVOT-VDC cue relationship is *asymmetric* between stop voicing categories. This observation may indicate a restriction on structured speaker variability for only those segments in a series (i.e., voiced and voiceless stops) that have some form of featural specification. It has been previously argued that Japanese is a 'voiced' language (Mester and Ito, 1989; Ito and Mester, 1995; Nasukawa, 2005) in being specified exclusively for a monovalent [voice] feature on voiced stops, with no featural specification for voiceless stops (e.g. Iverson and Salmons, 1995; Salmons, 2019). Furthermore, the lack of an observed correlation across cues may suggest that pVOT and VDC do not share an intrinsic link, potentially reflecting different articulatory pressures on their usage. This lack of a correlation, however, does not rule out a relationship between the cues: it is possible that VDC and pVOT, as measured here, simply do not capture the dimensions in which these cues may be related.

The within-cue findings (Section 2.4.1) suggest that speakers can use cues independently to mark a linguistic contrast *without* maintaining the same cross-category relationships across more than one phonetic cue. This supports a restricted form of structured variability, constraining the predictability of speakers of spontaneous Japanese in their realisation of phonological categories along a single phonetic dimension. Crucially, speakers use two cues to *separately* realise the same phonological contrast. In this sense, the structured variability is *constrained*: in this study, speaker variability is present within the use of a single acoustic cue, but speakers are less consistent in simultaneous use of multiple cues to the stop voicing contrast.

When considered from the perspective of a 'principle of uniformity' constraining phonetic variation (Chodroff and Wilson, 2017, 2018), our results provide some evidence for uniformity across speakers: namely, speakers are highly consistent *within* cues in signalling stop voicing contrasts. Our findings also demonstrate that a principle of uniformity is likely subject to constraints: here we find evidence of speakers covarying within individual cues, as opposed to covarying across more than one cue in marking the same contrast. Japanese differs from English in how the stop voicing contrast is specified: Japanese maintains a 'hybrid' stop voicing system involving the use of both positive VOT and voicing during closure (e.g. Nasukawa, 2005). Thus our evidence for covariation from Japanese stop voicing suggests that phonetic uniformity is constrained by language-specific properties. Our study emphasises the importance of examining the evidence for uniformity in a range of empirical contexts, and especially across languages which differ in their phonetic implementation of a given phonological contrast.

2.6 Conclusion

This study has examined stops in spontaneous Japanese and demonstrated that structured variability is present in a new empirical setting, and that it is constrained in ways not straightforwardly predicted from studies mainly focusing on English. Specifically, the constraint arises from the linguistic specification and phonetic implementation of stop voicing in Japanese which requires a different configuration of acoustic cues from English. Such a finding motivates an expanded search for structured speaker variability across more languages and phonetic cues. Within Japanese, for example, this could mean including F0 as an acoustic cue, given its increasing importance for the stop voicing contrast (Kong et al., 2014; Gao et al., 2019; Gao and Arai, 2019). Our study provides the first sketch for a more complex appreciation of how speaker variability is structured. It also motivates increasing the range of studies on structured variability across languages, cues, and contrasts (Bang, 2017; Hullebus et al., 2018; Hauser, 2019).

2.7 Appendices

2.7.1 Population-level effects (pVOT)

Table 2.3: Estimate ($\hat{\beta}$), error, and 95% credible intervals for all population-level ('fixed effect') predictors for log-transformed pVOT.

Predictor	$\hat{\beta}$	Error	2.5% CrI	97.5% CrI
Intercept	3.11	0.02	3.08	3.15
Voicing	-0.51	0.02	-0.54	-0.48
Gender	-0.09	0.03	-0.15	-0.03
Previous phoneme manner (long)	0.03	0.00	0.03	0.04
Previous phoneme manner (nasal)	0.03	0.00	0.02	0.04
Birth year (1960-69)	0.04	0.02	-0.01	0.09
Birth year (1950-59)	0.03	0.02	-0.02	0.08
Birth year (1940-49)	0.00	0.03	-0.06	0.06
Birth year (1930-39)	-0.02	0.04	-0.09	0.05
Place of articulation (alveolar)	-0.18	0.01	-0.20	-0.15
Place of artciulation (velar)	-0.12	0.01	-0.14	-0.10
Speech style (public speaking)	-0.10	0.00	-0.11	-0.09
Style style (dialogue)	0.01	0.00	0.00	0.02
Break Index (2)	0.05	0.00	0.05	0.06
Break Index (3)	0.05	0.00	0.04	0.05
Frequency (log)	-0.04	0.01	-0.05	-0.03
Speech rate (mean)	-0.06	0.03	-0.12	0.01
Speech rate (local)	-0.03	0.00	-0.04	-0.02
Preceding pause	0.04	0.01	0.02	0.05
Vowel height	0.14	0.01	0.11	0.16
Voicing : Gender	0.08	0.02	0.03	0.12
Voicing : Previous phoneme manner (long)	0.02	0.01	0.01	0.03
Voicing : Previous phoneme manner (nasal)	0.02	0.01	0.00	0.04
Voicing : Birth year (1960-69)	-0.03	0.02	-0.07	0.00
Voicing : Birth year (1950-59)	-0.05	0.02	-0.08	-0.01
Voicing : Birth year (1940-49)	0.04	0.02	0.00	0.08
Voicing : Birth year (1930-39)	-0.03	0.03	-0.08	0.02
Voicing : Place of articulation (alveolar)	0.05	0.02	0.02	0.09
Voicing : Place of articulation (velar)	0.06	0.01	0.04	0.09
Voicing : Speech style (public speaking)	0.03	0.01	0.01	0.04
Voicing : Speech style (dialogue)	-0.02	0.01	-0.03	0.00
Voicing : Break Index (2)	-0.06	0.01	-0.07	-0.05
Voicing : Break Index (3)	-0.04	0.00	-0.05	-0.03
Voicing : Frequency (log)	0.01	0.01	-0.01	0.04
Voicing : Speech rate (mean)	0.05	0.03	0.00	0.10
Voicing : Speech rate (local)	0.00	0.01	-0.02	0.01
Voicing : Preceding pause	-0.06	0.02	-0.10	-0.03
Voicing : Vowel height	-0.08	0.02	-0.13	-0.03

2.7.2 Population-level effects (VDC)

Table 2.4: Estimate ($\hat{\beta}$), error, and 95% credible intervals for all population-level ('fixed effect') predictors for VDC (logit-scale).

Predictor	$\hat{\beta}$	Error	2.5% CrI	97.5% CrI
Intercept	-1.13	0.12	-1.36	-0.90
Voicing	2.99	0.14	2.72	3.25
Gender	0.12	0.18	-0.23	0.48
Previous phoneme manner (long)	0.01	0.03	-0.06	0.07
Previous phoneme manner (nasal)	-0.17	0.05	-0.27	-0.08
Birth year (1960-69)	0.33	0.14	0.04	0.61
Birth year (1950-59)	0.40	0.15	0.10	0.69
Birth year (1940-49)	-0.01	0.18	-0.35	0.34
Birth year (1930-39)	-0.36	0.21	-0.77	0.06
Place of articulation (alveolar)	0.00	0.07	-0.14	0.13
Place of articulation (velar)	0.13	0.05	0.04	0.22
Speech style (public speaking)	0.13	0.04	0.04	0.21
Speech style (dialogue)	-0.42	0.05	-0.52	-0.33
Break Index (2)	0.39	0.03	0.32	0.45
Break Index (3)	0.53	0.02	0.49	0.58
Frequency (log)	0.17	0.04	0.09	0.26
Speech rate (mean)	-0.57	0.19	-0.95	-0.20
Speech rate (local)	-0.16	0.04	-0.23	-0.09
Preceding pause	-3.24	0.16	-3.56	-2.93
Vowel height	0.12	0.07	-0.02	0.26
Voicing : Gender	0.06	0.20	-0.34	0.45
Voicing : Previous phoneme manner (long)	-0.20	0.06	-0.32	-0.07
Voicing : Previous phoneme manner (nasal)	-0.09	0.07	-0.22	0.04
Voicing : Birth year (1960-69)	-0.03	0.17	-0.36	0.29
Voicing : Birth year (1950-59)	0.04	0.17	-0.30	0.38
Voicing : Birth year (1940-49)	0.05	0.20	-0.34	0.44
Voicing : Birth year (1930-39)	-0.32	0.24	-0.78	0.14
Voicing : Place of articulation (alveolar)	0.24	0.12	0.01	0.46
Voicing : Place of articulation (velar)	0.13	0.09	-0.04	0.31
Voicing : Speech style (public speaking)	0.52	0.07	0.38	0.67
Voicing : Speech style (dialogue)	0.13	0.08	-0.03	0.28
Voicing : Break Index (2)	-0.54	0.06	-0.64	-0.42
Voicing : Break Index (3)	-0.57	0.03	-0.63	-0.50
Voicing : Frequency (log)	0.14	0.08	-0.02	0.30
Voicing : Speech rate (mean)	0.11	0.22	-0.31	0.55
Voicing : Speech rate (local)	0.10	0.06	-0.02	0.22
Voicing : Preceding pause	2.00	0.21	1.58	2.41
Voicing : Vowel height	0.60	0.15	0.31	0.90

Preface to Chapter 3

Chapter 2 examined how speakers systematically vary in the use of two acoustic cues – VOT and closure voicing – in the realisation of the Japanese stop voicing contrast. It was found that, in spite of variability in the overall use of each cue, the relative use of a single cue was highly constrained across speakers. This close relationship across speakers was not observed *across* cues, however, suggesting that structured speaker variability may be highly language and cue-specific. These results were discussed in terms of previous research examining structured speaker variability, which have largely addressed stops in languages such as English and German, and considers that the language-specific implementation of the voicing contrast may play a role in what acoustic cues are constrained in their patterns of variation.

In order to further investigate whether variability is constrained with respect to the language-specific implementation of phonetic cues, Chapter 3 expands the analysis of structured variability by focusing on variation in the voicing contrast in a different context. Specifically, this chapter examines how the word-final voicing contrast – represented by durational differences in the preceding vowel – is structured across both speakers and dialects of English. Whilst this phenomenon has been extensively studied within the phonetic literature, little is known about the scope of variability across dialects and speakers of English: considering dialects as linguistic varieties with similar phonological structures, Chapter 3 explores the extent to which the primary cue to final consonant voicing exhibits structure in its patterns of variability.

Chapter 3

Towards 'English' phonetics: variability in the pre-consonantal voicing effect across English dialects and speakers

3.1 Introduction

There exist a large number of well-studied properties of speech that are known to vary across languages and communities of speakers, which have long been of interest to sociolinguists and phoneticians. One dimension of this variability, which is the focus of this study, is that of variation *within languages*: across dialects and their speakers. For example, the deletion of word-final /t/ and /d/ segments (in e.g., *mist*, *missed*) has been shown to vary across a wide range of dialects and speech communities (e.g. Labov et al., 1968; Guy, 1980; Tagliamonte and Temple, 2005), as have the dialect-specific realisation of English vowels (e.g. Thomas, 2001; Clopper et al., 2005; Labov et al., 2006), and variation in the degree of aspiration in English voiced and voiceless stops (e.g. Docherty, 1992; Sonderegger et al., 2017; Stuart-Smith et al., 2015). The study of this kind of variation provides a means of understanding the sources and structures of variability within languages: both in how particular dialects may systematically differ from each other, and how the variable realisation of speech sounds maps to speakers' cognitive representation of language and speech (Liberman et al., 1967; Lisker, 1985; Kleinschmidt, 2018). Despite decades of research, however, there is much we do not know about the scope, extent, and structure of this kind of language-internal variability. Within the phonetic literature, most research has focused on highly-controlled speech styles in 'laboratory settings', generally focusing on a single dialect in each study; much of the work focusing on phonetic variability in spontaneous speech is on single dialects (e.g. Ernestus et al., 2015). The sociolinguistic and dialectological literatures have often examined spontaneous speech, with some notable cross-dialectal studies (e.g. Labov et al., 2006; Clopper et al., 2005; Jacewicz and Fox, 2013), but nonetheless primarily focus on variation in vowel quality. Increasingly, however, research within phonetics and sociophonetics is being performed at a larger scale *across* speech communities (Labov et al., 2006, 2013; Yuan et al., 2006, 2007; Yuan and Liberman, 2014; Coleman et al., 2016; Liberman, 2018), driven by the development of new speech processing tools and data sharing agreements. This 'large-scale' approach is applied here to one such well-studied variable, the pre-consonantal voicing effect, as a means of characterising its degree and structure of variability in a single phonetic effect across English dialects and speakers.

The pre-consonantal voicing effect (henceforth *Voicing Effect*, VE) refers to vowels preceding voiced obstruents being consistently longer than their voiceless counterparts, such as the differences in *beat-bead* and *mace-maze* (House and Fairbanks, 1953; House, 1961). The VE has been reported – to greater or lesser extent – in a range of languages (Zimmerman and Sapon, 1958; Chen, 1970), though it varies in size based on properties of the phonetic environment, such as whether the obstruent is a stop or fricative, the height of the vowel, and many others (Klatt, 1973; Crystal and House, 1982; Port and Dalby, 1982). The evidence for the English VE to date is sourced predominantly from laboratory studies of highly-controlled speech, often in citation form, recorded from small numbers of often standard General American English speakers (e.g. Rositzke, 1939; House and Fairbanks, 1953; Peterson and Lehiste, 1960; House, 1961; Luce and Charles-Luce, 1985; Crystal and House, 1982). On the basis of this evidence, the VE has been noted for being particularly large in English relative to other languages (Zimmerman and Sapon, 1958; Chen, 1970), and has long been suggested as a prominent cue to consonant voicing in English (Denes, 1955; Klatt, 1973). This in turn has motivated claims that the VE is learned in English, as opposed to being a low-level phonetic property in other languages (Fromkin, 1977; Keating, 2006; Solé, 2007). At the same time, numerous questions about the nature and extent of the VE in English remain unexplored. In this study, we will examine the variability in the VE across a range of English dialects, focusing on the following two research questions: (1) *how large is the VE as realised in spontaneous English speech?*, and (2) *how much does the VE vary across dialects and speakers?* In addressing these questions, we hope to gain insight into a number of open issues, including the extent to which there is a single 'English' VE or whether dialects differ in the magnitude of the effect, as well as the range of VE sizes across individual speakers of a given dialect.

This paper answers these questions by taking a 'large-scale' approach to the study of the VE. Concretely, this refers to the use of a large amount of acoustic data, collected from a large number of speakers across a range of English dialects. This analysis falls within the framework of the *SPeech Across Dialects of English* (SPADE) project (Sonderegger et al., 2020b, https://spade.glasgow.ac.uk/), which aims to consider phonetic and phonological variation in British and North American English across time and space through the use of automated acoustic analysis of features across English dialects occurring in many corpora. The methodological and research goals of the SPADE project are exemplified through this study of the English VE, specifically by the use of multiple corpora of diverse sources and structures, and the use of linguistic and acoustic analysis via the *In-*

tegrated Speech Corpus ANalysis (ISCAN) tool (McAuliffe et al., 2019), developed as part of the broader SPADE project. Both the volume and complexity of the resulting data and the goals of the study motivate the need for appropriately-flexible approaches to the statistical analysis: specifically, the data is statistically analysed using Bayesian regression models (Carpenter et al., 2017), which enable us to accurately estimate the size of the VE across dialects and speakers directly, whilst controlling for the complex nature of the spontaneous speech data.

The structure of this paper is as follows. Section 3.2 outlines previous work on the VE, and some of the outstanding questions related to our current understanding of its variability. Section 3.3 describes the data: the corpora of different dialects from SPADE. Sections 3.4 and 3.5 describe the methodological approach: the process of acoustic and statistical analysis of the data. The results of this analysis are reported in Section 3.6, and then discussed with respect to our specific research questions in Section 3.7 and concluding in Section 3.8.

3.2 The voicing effect (VE)

The observation that vowels preceding voiced obstruents are consistently longer than before voiceless obstruents was first noted in early phonetics textbooks (e.g. Sweet, 1880; Kenyon, 1940; Thomas, 1947; Jones, 1948) and in preliminary experimental work from the first half of the twentieth century (Heffner, 1937; Rositzke, 1939; Hibbitt, 1948). Studies explicitly manipulating the VE in English observed an effect of around 1.45 – that is, vowels before voiced consonants were longer than before voiceless consonants by a ratio of around 2:3 (House and Fairbanks, 1953; House, 1961), and this effect was a cue to the voicing of the obstruent (Denes, 1955; Lisker, 1957; Raphael, 1972).

In these studies, VE was shown to be affected by consonant manner: namely, that fricatives showed a smaller or minimal VE compared to stops (Peterson and

Lehiste, 1960), and less-robustly cued the voicing of the final consonant (Raphael, 1972). Initial studies of connected speech suggested that the size of the VE in this type of speech is more variable: VEs in carrier sentences are similar to those in isolated words (Luce and Charles-Luce, 1985),[1] whilst vowels in read or spontaneous speech exhibit smaller VE sizes of around 1.2, and a negligible VE for fricatives (Crystal and House, 1982; Tauberer and Evanini, 2009). VE size is also modulated by the overall length of the vowel, which is hypothesised to be due to an intrinsic incompressibility of the vowel, limited by the minimal time required to perform the articulatory motor commands necessary for vowel production (Klatt, 1976). This general suggestion has been supported by observations that VE is smaller for unstressed and phrase-medial vowels (Umeda, 1975; Klatt, 1976), and vowels produced at a faster speech rate (Crystal and House, 1982; Cuartero, 2002). The VE is thus modulated by a range of phonetic factors, and largely predict a reduction of VE size in instances where vowels are generally shorter; vowels that undergo 'temporal compression' have a reduced capacity to maintain a large VE size, and so VE is minimised. As these effects have only been investigated in laboratory speech, it is not clear whether the size and direction of these effects are maintained in less-controlled spontaneous speech styles.

Examining the VE across languages, Zimmerman and Sapon (1958) first observed that whilst English speakers produced a robust VE, Spanish speakers did not modulate vowel length in the same way, though this study did not control for the syllabic structure of test items. Comparing across English, French, Russian, and Korean, Chen (1970) observed that all four languages produced a VE size of at least 1.1, though all languages had different VE sizes (English = 1.63, French = 1.15, Russian = 1.22, Korean = 1.31). This was interpreted as evidence that VE is a phonetically-driven effect with additional language-specific phonological specification (Fromkin, 1977). Mack (1982), comparing English and French monolin-

[1] Harris and Umeda (1974), in their study of overall vowel duration, attribute this difference to a 'mechanical' prosody as a consequence of numerous repetitions.

guals with bilinguals, observed that English monolinguals maintained a substantially larger VE than French monolinguals, whilst the French-English bilinguals also produced the shorter French-style pattern instead of adapting to the larger English VE pattern. Keating (1985) suggested that VE is 'phonetically-preferred', though ultimately controlled by the grammar of the particular language. English, then, is expected to have a larger VE than other languages, though it is not known if the English VE is of a comparable size in spontaneous speech.

The work discussed above has not differentiated between varieties of English, and cross-linguistic comparisons of VE have presumed that a single 'English' VE size exists. Little work has focused on variation in VE across English dialects beyond a small number of studies on specific dialects. One dialect group of interest has been Scottish Englishes and the application of the Scottish Vowel Length Rule (SVLR), where vowels preceding voiced fricatives and morpheme boundaries are lengthened, whilst all other contexts have short vowels (Aitken, 1981), and hence do not show the VE. In studies of the SVLR, some East Coast Scotland speakers show some evidence of the VE in production (Hewlett et al., 1999), whilst VE-like patterns were not observed in spontaneous Glaswegian (Rathcke and Stuart-Smith, 2016). On the other hand, studies of African American English (AAE) have claimed that voiced stops undergo categorical devoicing in this variety, which has resulted in additional vowel lengthening before voiced stops to maintain the pre-consonantal voicing contrast (Holt et al., 2016; Farrington, 2018). Only one study has previously compared the VE *across* English dialects in spontaneous speech. Tauberer and Evanini (2009), using interview data from the *Atlas of North American English* (Labov et al., 2006), observe that North American English dialects vary in their VE values, ranging from 1.02 to 1.33, and that dialects with shorter vowels on average (New York City) also show a smaller-than-average VE size (1.13). Moreover, despite recognition that individual speakers may exhibit variability in their VE sizes (Rositzke, 1939; Summers, 1987), no study has formally examined the extent of variability across speakers, nor how dialects may differ in the degree of VE variability amongst its speakers. The two patterns observed for Scottish and African American English suggest that English dialects can maintain relatively 'small' (or no), and 'large' VEs respectively; we know little about the degree of VE variability beyond these dialects without a controlled study across multiple English varieties, which is one of the goals of this study.

Whilst a large number of studies on the VE have provided useful information for its realisation in English and other languages, there are still a range of outstanding questions that can be addressed through a large-scale cross-dialectal approach. To what extent is the VE a *learned* property of a given language, compared with an *automatic* consequence of low-level phonetic structure? Much of the discussion with respect to variation in VE has revolved around differences across *languages* (Chen, 1970; Keating, 1985), which may differ both in their phonetic realisation of segments but also the phonological representation of those segments. In this sense, examining VE variability internal to a language (i.e., across *dialects*) potentially avoids this problem; the specification of phonological categories – here, the voicing status of final obstruents – is expected be largely consistent within a language, meaning that language-internal variability may be driven by only differences in phonetic implementation.

Little is known about how English dialects may vary in their implementation of the VE, and so a range of possibilities exist for how dialects might compare. One possibility is that, with the exception of varieties with specific phonological rules interacting with the VE, dialects might cluster around a single 'English' VE value, potentially of the size reported in the previous literature. Such a finding would support the previous approach in the literature, in terms of English compared to other languages, and suggest that dialects do not differ in how the final voicing contrast is phonetically implemented. Alternatively, dialects may differ gradiently from each other, and so may show a continuum of possible dialect-specific VE sizes. If dialects do differ in their VE size in this way, this would suggest that the previous literature on the VE in 'English' accounts for just a fraction of the possible VE realisations across English, and would provide evidence that individual English dialects differ in their phonetic implementation of an otherwise 'phonological' contrast (Keating, 1984, 1985).

Similarly, little is known about how individual speakers vary in the VE, and what the overall distribution of speaker VE sizes is. Synchronic variability across speakers is one of the key inputs to sound change (Ohala, 1989; Baker et al., 2011), and also defines the limits of a speech community, i.e., speakers who share sociolinguistic norms in terms of production and social evaluation (e.g. Labov, 1972). Whilst dialects may differ in the realisation of segments or the application of phonological processes, dialect-internal variability is potentially more limited if a phonetic alternation such as the VE is critical to speech community membership.

3.3 Data for this study

The varieties of English included in this study are from North America, Great Britain, and Ireland. For the purposes of this study, North American dialects refer to the regions of the United States and Canada outlined in *The Atlas of North American English*, which is based around phonetic, not lexical, differences between geographic regions (Labov et al., 2006; Boberg, 2018). For Canadian data specifically, the primary distinction was made between 'urban' and 'rural' speakers, based on its relative importance noted in comparison to much weaker geographic distinctions, at least for the corpus which makes up most Canadian data in this study (Rosen and Skriver, 2015). Within the British and Irish groups, dialects from England in this study are defined in terms of Trudgill's dialectal groupings (Trudgill, 1999), which groups regions in terms of both phonological and lexical similarity. Due to the lack of geographical metadata for speakers from Ireland and Wales, these dialects were simply coded as 'Ireland' and 'Wales' directly. Scottish Englishes are grouped based on information from *The Scottish National Dictionary*.[2] The data used in this study comes from the SPADE project, which aims to bring together and analyse over 40 speech corpora covering English speech across North America, the United Kingdom, and Ireland. In this study, we analyse data from 15 of these corpora, which together cover 30 different English dialects from these regions, comprised of speech from interviews, conversations, and reading passages. A basic description of each of these corpora is given below, outlining the type of speech and phonetic alignment tools used.

- *Audio British National Corpus* (AudioBNC, Coleman et al., 2012): The spoken sections of the British National Corpus, originally containing speech from over 1,000 speakers. However, due to a range of recording issues (e.g., overlapping speech, background noise, microphone interference), a large portion of the corpus is inaccurately aligned. In order to define a subset of the AudioBNC which maximises the accuracy of the alignment, utterances were kept if they met a number of criteria: the utterance length was greater than one second, that the utterance contained at least two words, that the mean harmonics-to-noise ratio of the recording was at least 5.6, and that the mean difference in segmental boundaries between the alignment and a re-alignment with the Montreal Forced Aligner (MFA, McAuliffe et al., 2017a) was at most 30ms.[3] 50 TextGrids from the remaining data were manually checked and deemed to be as approximately accurate as that of normal forced-alignment.

[2] Part of *The Dictionary of the Scots Language* (https://dsl.ac.uk/).

[3] We are grateful to Michael Goodale for designing and performing this filtering protocol.

- *Brains in Dialogue* (Solanki, 2017): recordings of 24 female Glaswegian speakers producing spontaneous speech in a laboratory setting. There are 12 recordings for each speaker, which were aligned with LaBB-CAT (Fromont and Hay, 2012).

- *Buckeye* (Pitt et al., 2007): spontaneous interview speech of 40 speakers from Columbus Ohio, recorded in 1990s-2000s. The Buckeye corpus is hand-corrected with phonetic transcription labels: these were converted back to phonological transcriptions in order to be comparable with data from the other corpora.

- *Corpus of Regional African American Language* (CORAAL, Kendall and Farrington, 2018): spontaneous sociolinguistic interviews with 100 AAE speakers from Washington DC, Rochester NY, and Princeville NC, recorded between 1968 and 2016, and aligned with the MFA.

- *Doubletalk* (Geng et al., 2013): recordings of paired speakers carrying out a variety of tasks in order to elicit a range of styles/registers in a discourse/interactive situation. Ten speakers make up five pairs where one member is a speaker of Southern Standard British English and the other member is a speaker of Scottish English.

- *Hastings* (Holmes-Elliott, 2015): recordings of sociolinguistic interviews with 46 speakers from Hastings in the south east of England, male and female, aged from 8-90, aligned using FAVE (Rosenfelder et al., 2014).

- *International Corpus of English – Canada* (ICE-Canada, Greenbaum and Nelson, 1996): interview and broadcast speech of Canadian English, recorded in the 1990s across Canada, and aligned using the MFA. Speaker dialect was defined in terms of their city or town of origin. In this study, we coded a speaker as ‘urban’ if their birthplace was a large Canadian city.

- *Canadian Prairies* (Rosen and Skriver, 2015): Spontaneous sociolinguistic interviews, recorded between 2010 and 2016, with speakers of varying ethnic backgrounds from the provinces of Alberta and Manitoba, conducted as part of the Language in the Prairies project, and was aligned using the MFA.

- *Modern RP* (Fabricius, 2000): reading passages by Cambridge University students recorded in 1990s and 2000s. The speakers were chosen for having upper middle-class backgrounds as defined by at least one parent having a professional occupation along with the speaker also having attended private schooling. The data used in this study come from a reading passage aligned with FAVE.

- *Philadelphia Neighborhood Corpus* (PNC, Labov and Rosenfelder, 2011a): sociolinguistic interviews with 419 speakers from Philadelphia, recorded between 1973 and 2013, and were aligned with FAVE.

- *Raleigh* (Dodsworth and Kohn, 2012): semi-structured sociolinguistic interviews of 59 White English speakers in Raleigh, North Carolina, born between 1955 and 1989, and aligned with the MFA.

- *Santa Barbara* (Bois et al., 2000): spontaneous US English speech, recorded in the 1990s and 2000s, from a range of speakers of different regions, genders, ages, and social backgrounds.

- *The Scottish Corpus of Texts and Speech* (SCOTS, Anderson et al., 2007): approximately 1300 written and spoken texts (23% spoken), ranging from informal conversations, interviews, etc. Most spoken texts were recorded since 2000.

- *Sounds of the City* (SOTC, Stuart-Smith et al., 2017): vernacular and standard Glaswegian from 142 speakers over 4 decades (1970s-2000s), collected from

Table 3.1: Number of speakers and tokens per dialect (left), and by corpora from which each dialect was derived.

Region	Dialect	*n* speakers	*n* tokens	Corpus	*n* speakers	*n* tokens
North America	Canada (rural)	52	9313	Canadian Prairies	44	8316
				ICE-Canada	8	997
	Canada (urban)	64	12124	Canadian Prairies	56	11939
				ICE-Canada	8	185
	Midwest US	40	5567	Buckeye	40	5567
	New England	24	1336	Santa Barbara	7	174
				Switchboard	17	1162
	North Midland US	46	3084	Switchboard	46	3084
	Northern Cities US	21	1377	Santa Barbara	21	1377
	Northern US	58	3086	Switchboard	58	3086
	NYC	25	1477	Santa Barbara	6	158
				Switchboard	19	1319
	Philadelphia	371	59581	PNC	371	59581
	Princeville NC (AAE)	71	6759	CORAAL	17	6759
	Raleigh US	92	3282	Raleigh	92	3282
	Rochester NY (AAE)	14	6308	CORAAL	14	6308
	South Midland US	108	8188	Switchboard	108	8188
	Southern US	44	2738	Santa Barbara	6	345
				Switchboard	38	2393
	Washington DC (AAE)	50	21205	CORAAL	50	21205
	Western US	100	5456	Santa Barbara	50	2900
				Switchboard	50	2556
United Kingdom & Ireland	Central Scotland	24	2426	SCOTS	24	2426
	East Central England	51	2544	Audio BNC	51	2544
	East England	229	20727	Audio BNC	132	6622
				Doubletalk	5	726
				Hastings	44	12642
				ModernRP	48	737
	Edinburgh	18	1148	SCOTS	18	1148
	Glasgow	177	33938	Brains in Dialogue	23	9210
				SCOTS	27	2294
				SOTC	127	22434
	Insular Scotland	8	351	SCOTS	8	351
	Ireland	19	624	Audio BNC	19	624
	Lower North England	60	3325	Audio BNC	60	3325
	North East England	17	488	Audio BNC	17	488
	Northern Scotland & Islands	33	2280	SCOTS	33	2280
	Scotland	70	3468	Audio BNC	65	2633
				Doubletalk	5	835
	South West England	50	2067	Audio BNC	50	2067
	Wales	41	2524	Audio BNC	41	2524
	West Central England	41	2615	Audio BNC	41	2615
Total		1964	229406			

historical archives and sociolinguistic surveys, aligned using LaBB-CAT.

- *Switchboard* (Godfrey et al., 1992): 2,400 spontaneous telephone conversations between random participants from the multiple dialect regions in the United States on a variety of topics, containing data from around 500 speakers.

The goals of this study are to examine the size and variability in the English VE in spontaneous speech, and in variation in the VE across dialects and individual speakers. Specifically, the kind of dialectal variability being addressed in this study is that of *regional* variability: variability by race or ethnicity is not being directly considered in this study, with the exception of three African American English varieties, given the particular observations about AAE with respect to the VE (Holt et al., 2016; Farrington, 2018). This study also does not focus on differences according to age, either age-grading or apparent/real-time change in the VE over time; only speech data recorded since 1990s was included; the other data recorded prior to 1990 was excluded from further analysis. Analysis of the role of age and time in the VE in these English dialects remains a subject for future study.

3.4 Data analysis

Having collected and organised the speech data into dialects, it is then possible to extract and acoustically analyse the data in the study: that is, going from raw data (audio and transcription files) to datasets which can be statistically analysed. As the corpora differ in their formats – the phone labels used, organisation of speaker data, etc – modifying the acoustic analysis procedure for each different corpus format would be both labour and time-intensive, as well as increase the risk that the analysis itself differed across corpora. In order to standardise the acoustic analysis across corpora, the *Integrated Speech Corpus ANalysis* (ISCAN) tool was developed for use in this kind of cross-dialectal study in the context of the SPADE project. This section provides a brief overview of the ISCAN system: see McAuliffe et al. (2017b, 2019) and the ISCAN documentation page[4] for details of the implementation.

The process of deriving a dataset from raw corpus files consists of three major steps. In the first step, individual speech corpora (in the form of sets of audio-transcription pairs) are *imported* into a graph database format, where each tran-

[4] https://iscan.readthedocs.io/.

sciption file is minimally composed of word and phone boundaries (e.g., word-level and phone-level tiers in a TextGrid), and these word-phone relationships are structurally-defined in the database (i.e., that each phone belongs to a word). Importers have been developed for a range of standard automatic aligners, including all formats of corpora described in Section 3.3. Corpora, represented in database format, can then be further *enriched* with additional structure, measurements, and linguistic information. For example, utterances can be defined as groups of words (separated by silence of a specified length, e.g., 150ms), syllables can be defined as a property between groups of adjacent phones. Once the database has been enriched with utterance and syllable information, speech rate (often defined as syllables per second within an utterance) can be calculated and included in the database. Similarly, information about words (such as frequency) or speakers (such as gender, age, dialect etc) can be added to the corpus from metadata files. Once a corpus has been sufficiently enriched with linguistic and acoustic information, it is then possible to perform a *query* on the corpus at a given level of analysis. This level of analysis refers to the level of the hierarchy on which the resulting datafile should use as the main level of observation, for example individual phones, syllables, or utterances. Filters can be applied to a query to restrict it to the particular contexts of interest, for example, including only syllables occurring at the right edge of an utterance, or vowels followed by a specific subset of phone types (e.g., obstruents). Finally, the resulting query can then be *exported* into a data format (currently CSV only) for further analysis.

Each corpus was processed using the ISCAN software pipeline, and then combined into a single 'master' dataset, containing all phonetic, dialect, and speaker information from all of the analysed corpora necessary to carry out the analysis of the VE below. As the vowel duration annotations from the corpora (except for Buckeye) were created via forced alignment with a minimum duration of 10ms and a time-step of 30ms, any token with a vowel duration below 50ms was excluded from further study, as is common in acoustic studies of vowel formants to exclude heavily reduced vowels (Dodsworth, 2013; Fruehwald, 2013). To reduce the additional prosodic and stress effects on vowel duration, the study only included vowels from monosyllabic words occurring phrase-finally, where a phrase is defined as a chunk of speech separated by 150ms of silence. Raw speech rate was calculated as syllables per second within a phrase, from which two separate speech rates were derived. First, a mean speech rate for each speaker was calculated, which reflects whether a speaker is a 'fast' or 'slow' speaker overall. From that mean speech rate, a local speech rate was calculated as the raw rate for the utterance subtracted from the given speaker's mean. This local speech rate can be interpreted as how fast or slow that speaker produced the vowel within that particular phrase *relative* to their average speech rate (Sonderegger et al., 2017; Cohen Priva and Gleason, 2018). Word frequency was defined using the SUBTLEX-US dataset (Brysbaert and New, 2009). The final dataset contained 229,406 vowel tokens (1,485 word types) from 1,964 speakers from 30 English dialects. Table 3.4 shows the number of speakers and tokens for each dialect, and how many speakers/tokens were derived from each speech corpus.

3.5 Statistical analysis

The research goals of this study focus on the size and variability of the VE in English spontaneous speech, and how the VE varies across dialects and speakers. These goals motivate an approach of *estimating* the size of the VE in these contexts, rather than testing whether the VE 'exists' or not. Whilst controlled laboratory experiments are explicitly designed to balance across these contexts (by including matching numbers of tokens with stops vs fricatives, using words with similar frequency, etc), spontaneous speech taken from corpora is rarely balanced in this sense: some speakers speak more than others, have different conversations leading to some combinations of segments occurring infrequently relative to others, speakers manage properties of their speech (such as speech rate) for communicative purposes which are generally absent in laboratory studies. In trying to obtain an accurate estimate of the VE (or indeed any other linguistic property), the unbalanced nature of spontaneous speech motivates the need for a statistical approach where individual factors of interest (e.g., obstruent manner of articulation, dialects, etc) can be explored whilst controlling for the influence of other effects. This approach – the use of multiple regression to model corpus data – is now common in phonetics and sociolinguistic research (e.g. Tagliamonte and Baayen, 2012; Roettger et al., 2019), but has not, to our knowledge, been used to analyse multiple levels of variability in the VE.

In this study, this approach to estimation is performed using Bayesian regression modelling. Whilst other multifactorial statistical models would also be valid, Bayesian models provide us with some advantages that make the goal of estimating the size of the VE easier. Mixed-models are ideal for use in this study, as these capture variability at multiple levels (the VE overall, across dialects, across speakers) and this variability is of direct interest for our research questions. Bayesian mixed models resemble more traditional linear mixed-effects (LME) models commonly used in linguistic and phonetic research, such as those performed with the *lme4* package (Bates et al., 2015), though differ in a few key respects. First, Bayesian models make it easy to calculate the *range* of possible VE sizes in each context, as opposed to a single value that would be output in LME models: whilst LME models provide ranges for 'fixed' effects (across all dialects/speakers), Bayesian models provide a range of possible sizes for each level (i.e., an individual dialect). In a Bayesian model, all parameters (coefficients) in the model are assumed to have a *prior* distribution of possible values, reflecting which effect sizes are believed to be more or less likely, before examining the data itself. The output of a Bayesian model is a set of *posterior* distributions, which result from combining the priors and the likelihood of observing the data. Each model parameter has its own posterior distribution, which each represent the range of values for that parameter that is consistent with both the modelled data, conditioned on prior expectations about likely values, and the structure of the model itself. Bayesian models are well-suited to the task in this study, as they allow for flexible fitting of model parameters, and allow the complex random-effects structures which are often recommended for fitting statistically-conservative models (Barr et al., 2013), but which often fail to converge in LME models (Nicenboim and Vasishth, 2016). See Vasishth et al. (2018b) for an introduction to Bayesian modelling applied to phonetic research.

A Bayesian mixed model of log-transformed vowel duration was fit using *brms* (Bürkner, 2018): a R-based front-end for the Stan programming language (Carpenter et al., 2017), containing the following population-level ('fixed effects') predictors: the **voicing** and **manner** of the following obstruent, vowel **height** (high vs non-high), the lexical **class** of the word (lexical vs functional), both **mean** and **local** speech rates, and lexical **frequency**. To observe how compression of the vowel influences VE size, interactions between all of these factors with obstruent voicing were also included. The continuous predictors (both speech rates, frequency), were centred and divided by two standard deviations (Gelman and Hill, 2007). The two-level factors (obstruent voicing, manner, vowel height, lexical class) were converted into binary (0,1) values and then centred.

The group-level ('random effects') structure of the model contained the complete set of model predictors for both dialects and speakers, nested within dialects. These terms capture two kinds of variability in the VE size: for each individual dialect, as well as the degree of variability across speakers – the nesting of speaker term inside dialects can be interpreted as capturing the variability in the size of the VE across speakers *within* a given dialect. Given the expectation that both the overall vowel duration (represented by the intercept) and the manner of the obstruent would affect the size of the VE, correlation terms between the intercept and both the consonant voicing and manner predictors, as well as for the interaction *between* the voicing and manner predictors, were included for both dialects and speakers. Random intercepts were included for words and phoneme labels, also nested within dialects. The model was fit using 8000 samples across 4 Markov chains (2000/2000 warmup/sample split per chain) and was fit with weakly informative 'regularising' priors (Nicenboim and Vasishth, 2016; Vasishth et al., 2018b): the intercept prior used a normal distribution with a mean of 0 and a standard deviation of 1 (written as $Normal(0, 1)$); the other fixed effects parameters used $Normal(0, 0.5)$ priors, with the exception of the obstruent voicing parameter which used a $Normal(0.1, 0.2)$ prior.[5] The group-level (for dialects, speakers) parameters used the *brms* default prior of a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 10. The correlations between group-level effects used the LKJ (Lewandowski et al., 2009) with ζ = 2, which gives lower prior probability to perfect (-1/1) correlations, as recommended by Vasishth et al. (2018b).

3.6 Results

The results in this study will be reported in the context of the two main research questions concerning VE variability (1) in spontaneous speech, and (2) across English dialects and individual speakers. The results are reported for each effect in terms of the median value with 95% credible intervals (CrIs), and the probability of that effect's direction. These values enable us to understand the *size*

[5] The values chosen for the obstruent voicing parameter reflect the decision to allow a wide range of possible VE sizes, including values both above and below those reported in the previous literature. A sensitivity analysis was performed using an additional model fit with a 'uniform' flat prior for the obstruent voicing parameter, which returned VE values differing by an order of 10^{-3}, suggesting that the decision for the weakly-informative prior did not adversely affect the reported results.

of the effect (i.e., the change in vowel duration) and the confidence in the effect's predicted direction. The strength of evidence for an effect is distinct from the strength of the effect itself: to evaluate the strength of evidence for an effect, we follow the recommendations of Nicenboim and Vasishth (2016) and consider there to be *strong* evidence of an effect if the 95% credible interval does not include 0, and *weak* evidence for an effect if 0 is within the 95% CrI but the probability of the effect's direction is at least 95% (i.e., that there is less than 5% probability that the effect changes direction). Evaluating the strength of an effect is determined with respect to effect sizes previously reported for laboratory (e.g. House and Fairbanks, 1953; House, 1961) and connected speech (Crystal and House, 1982; Tauberer and Evanini, 2009). The degree of variability across dialects can be compared with the findings of Tauberer and Evanini (2009); as there is no known comparison for speaker variability, this will be compared to variability across dialects as an initial benchmark.

3.6.1 The voicing effect in spontaneous speech

Table 3.2 reports the population-level ('fixed') effects for each parameter in the fitted model. The 'overall' VE size averaging across dialects, which is between 1.09 and 1.2, is estimated to be smaller than reported in previous laboratory studies ($\hat{\beta}$ = 0.14, CrI = [0.09, 0.19], Pr($\hat{\beta} > 0$) = 1)[6] and more consistent with VE sizes reported in studies of connected and spontaneous speech (Crystal and House, 1982; Tauberer and Evanini, 2009).

Looking at how the overall VE size for all dialects is modulated by phonetic context, there is weak evidence that the manner of the following obstruent modulates VE size ($\hat{\beta}$ = -0.04, CrI = [-0.10, 0.02], Pr($\hat{\beta} < 0$) = 0.91): whilst stops appear to have a larger VE size (Figure 3.1, top left), the uncertainty in VE size

[6] As vowel duration was log-transformed prior to fitting, effects are interpreted by taking the exponent of the model parameter's value, e.g., $e^{0.19} = 1.2$, which refers to a vowel duration increase of 20%.

Table 3.2: Posterior mean ($\hat{\beta}$), estimated error, upper & lower credible intervals, and posterior probability of the direction of each population-level parameter included in the model of log-transformed vowel duration.

Parameter	$\hat{\beta}$	Est.Error	95% CrI	Pr($\hat{\beta}$ <>0)
Intercept	-1.99	0.02	[-2.03, -1.96]	1
Obstruent voicing	0.14	0.03	[0.09, 0.19]	1
Obstruent manner	0.05	0.02	[0.02, 0.08]	1
Vowel height	-0.22	0.02	[-0.25, -0.18]	1
Lexical class	-0.14	0.03	[-0.21, -0.08]	1
Speech rate (mean)	-0.22	0.01	[-0.24, -0.20]	1
Speech rate (local)	-0.28	0.01	[-0.30, -0.26]	1
Lexical frequency	-0.05	0.01	[-0.08, -0.03]	1
Voicing : Manner	-0.04	0.03	[-0.10, 0.02]	0.91
Voicing : Height	0.07	0.02	[0.02, 0.11]	1
Voicing : Class	-0.07	0.03	[-0.13, 0.00]	0.97
Voicing : Mean rate	-0.01	0.01	[-0.03, 0.01]	0.77
Voicing : Local rate	-0.06	0.01	[-0.08, -0.03]	1
Voicing : Frequency	-0.07	0.02	[-0.11, -0.03]	1

for each obstruent manner (represented by the spread of the credible intervals) suggests that it is possible there is no difference in VE size between both obstruent manners. Whilst high vowels are shown to be shorter than non-high vowels overall ($\hat{\beta}$ = -0.22, CrI = [-0.25, -0.18], Pr($\hat{\beta} < 0$) = 1), there is strong evidence that high vowels have a larger VE than non-high vowels ($\hat{\beta}$ = 0.07, CrI = [0.02, 0.11], Pr($\hat{\beta} > 0$) = 1). There is a similarly strong effect for lexical class ($\hat{\beta}$ = -0.07, CrI = [-0.13, 0.00], Pr($\hat{\beta} < 0$) = 0.97), where functional words have smaller VEs than open-class lexical items (Figure 3.1, top right). Lexical frequency also has a strong and evident effect on VE size ($\hat{\beta}$ = -0.07, CrI = [-0.11, -0.03], Pr($\hat{\beta} < 0$) = 1), where higher-frequency words have smaller VEs than their lower-frequency counterparts (Figure 3.1, bottom left), whilst local speech rate also reduces VE size ($\hat{\beta}$ = -0.06, CrI = [-0.08, -0.03], Pr($\hat{\beta} < 0$) = 1; Figure 3.1, bottom middle). For mean speaking rate, however, the effect on VE is both small with weak evidence ($\hat{\beta}$ = -0.01, CrI = [-0.03, 0.01], Pr($\hat{\beta} < 0$) = 0.77): this is reflected in Figure 3.1 (bottom right), where the difference between faster and slower speakers has

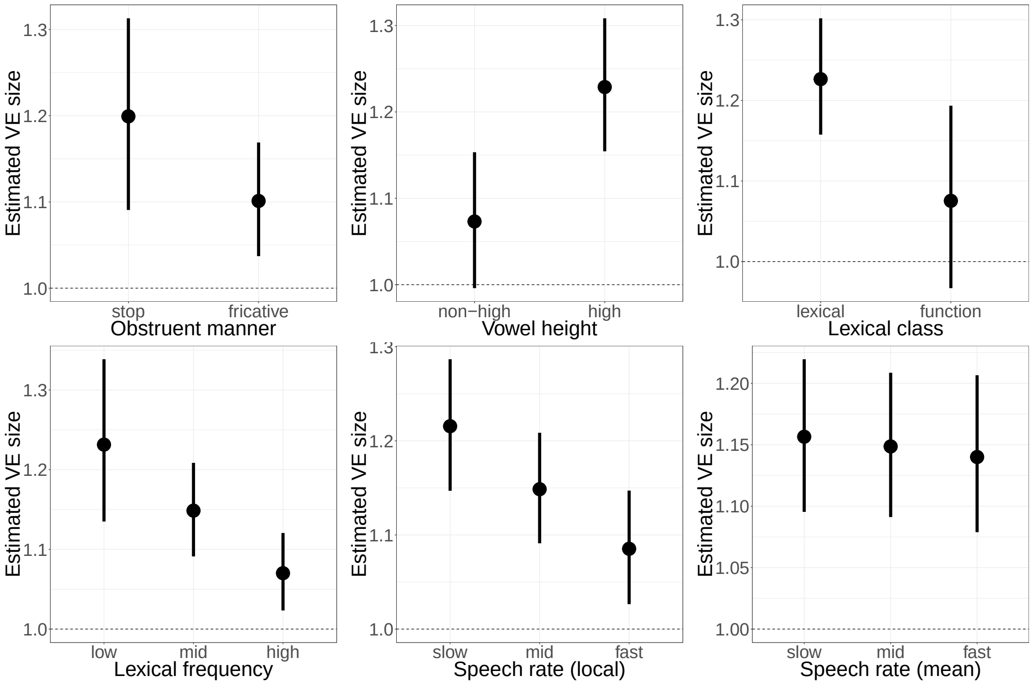

Figure 3.1: Modulation of VE size in different phonetic contexts: obstruent manner (top left), vowel height (top middle), lexical class (top right), frequency (bottom left), local (bottom middle) and mean (bottom right) speech rates. Points and error bars indicate the posterior mean value with 95% credible intervals, whilst holding all other predictors at their average values. Dashed line indicates no difference between vowels preceding voiced or voiceless consonants. For continuous predictors (frequency, speech rates), the estimate VE size is shown at three values for clarity.

a negligible effect on VE size. These results generally suggest that shorter vowels (within-speaker) tend to have smaller VE sizes, consistent with the temporal compression account (Klatt, 1973): the apparent exception to this is the relationship between VE size and vowel height, which is addressed in Section 3.7.

3.6.2 Voicing effect across dialects and speakers

Turning to dialectal variability in VE, we observe that the dialect variation in VE (the dialect-level standard deviation, $\hat{\sigma}_{dialect}$) is between 0.07 and 0.12: this can be interpreted as meaning that the difference in VE size between a 'low' and

Table 3.3: Posterior mean ($\hat{\sigma}$), estimated error, and 95% credible intervals for dialect and speaker-level parameters related to obstruent voicing included in the model of log-transformed vowel duration.

Level	Parameter	$\hat{\sigma}$	Est.Error	95% CrI
Dialect	Intercept	0.05	0.01	[0.03, 0.07]
	Obstruent Voicing	0.09	0.01	[0.07, 0.12]
	Voicing : Manner	0.12	0.02	[0.09, 0.16]
	Voicing : Height	0.04	0.01	[0.01, 0.06]
	Voicing : Class	0.06	0.01	[0.04, 0.09]
	Voicing : Mean Rate	0.02	0.01	[0.00, 0.05]
	Voicing : Local Rate	0.05	0.01	[0.03, 0.07]
Speaker	Intercept	0.10	0.00	[0.09, 0.10]
	Obstruent Voicing	0.08	0.00	[0.07, 0.08]
	Voicing : Height	0.11	0.01	[0.10, 0.12]
	Voicing : Manner	0.11	0.01	[0.10 0.13]
	Voicing : Class	0.13	0.01	[0.11, 0.14]
	Voicing : Local Rate	0.09	0.01	[0.08, 0.11]

'high' VE dialect is between 32% and 61%.[7] This is comparable with the range of possible values for the overall VE (between 0.09 and 0.19, Table 3.2 row 2). To understand whether this constitutes a 'large' degree of variability, one metric is to assess whether a 'low VE' dialect would actually have a reversed effect direction (voiceless > voiced), which is tested by subtracting 2 x $\hat{\sigma}_{dialect}$ from the overall VE size and comparing to 0. There is little evidence that dialects differ enough to change direction ($\hat{\beta}$ = -0.05, CrI = [-0.09, 0], Pr($\hat{\beta} > 0$) = 0.06), which suggests that whilst individual dialects differ in the *size* of the VE, no dialect fully differs in the *direction* of the effect (i.e., no dialect's credible interval is fully negative).

Another way of understanding the degree of dialectal variability in VE is to examine the predicted VE for individual dialects. As shown in Figure 3.2, dialects appear to differ gradiently from each other, ranging from dialects with effectively-null VE to those with strong evidence for large VEs. The Scottish dialects of Central Scotland and Edinburgh have VEs of at most 1.06 and 1.09

[7] The value is multiplied by 4 to get the 95% range of values = $2\hat{\sigma}_{dialect}$ for both sides of the distribution = 0.28, which is then back-transformed from log via the exponential function = $e^{0.28}$ = 1.32.

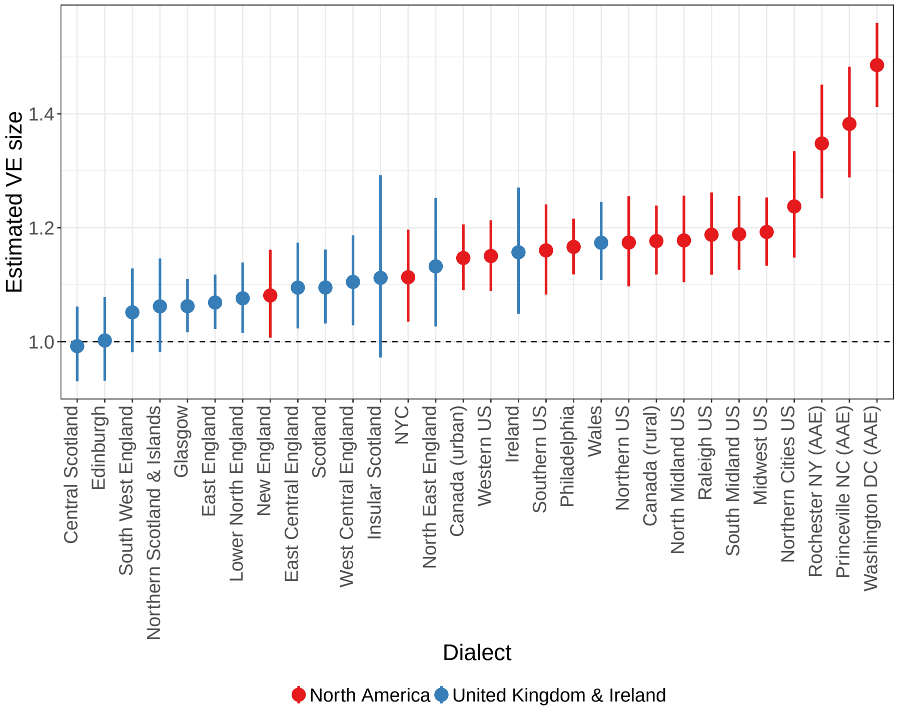

Figure 3.2: Estimated VE size for each dialect analysed in this study (red = North American, blue = United Kingdom & Ireland). Points and errorbars indicate the posterior mean value with 95% credible intervals, whilst holding all other predictors at their average values. Dashed line indicates no difference between vowels preceding voiced or voiceless consonants.

respectively, based on their upper credible interval value, whilst their median values (indicated by the points in Figure 3.2) indicate that the most likely VE size is around 0 (Central Scotland: $\hat{\beta} = 0.99$, CrI = [0.93, 1.06]; Edinburgh: $\hat{\beta} = 1.01$, CrI = [0.93, 1.09]): indeed, all Scottish dialects have a predicted VE size of 1.16 at the highest, with most of these having median values less than 1.1 (Table 3.4). North American dialects, in contrast, all have robustly positive VE values (no credible interval crosses the 0 line) and are generally larger than the British and Irish variants, shown by the position of red (North American) and blue (United Kingdom & Ireland) points respectively in Figure 3.2. In particular, the AAE dialects have the largest VEs in the sample, which are all robustly larger than the average 'English' VE size (Rochester NY: $\hat{\beta} = 1.35$, CrI = [1.27, 1.44]; Princeville NC: $\hat{\beta} = 1.39$, CrI = [1.31, 1.48]; Washington DC: $\hat{\beta} = 1.49$, CrI = [1.42, 1.56]): this is consis-

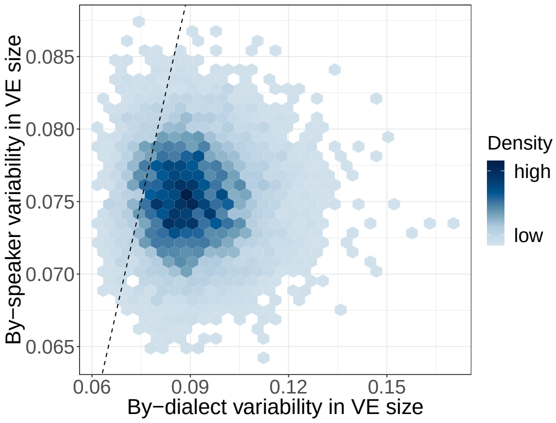

Figure 3.3: Heatmap of posterior samples of by-dialect ($\hat{\sigma}_{dialect}$) and by-speaker ($\hat{\sigma}_{speaker}$) voicing effect standard deviations. Equal variability is indicated by the dashed line, with darker shades indicating a greater density of samples.

tent with previous studies of studies on AAE, which posit that final devoicing of word-final voiced obstruents results in compensatory vowel lengthening (Holt et al., 2016; Farrington, 2018).

Turning to variability in VE across individual speakers, we observe that speakers are estimated to vary within-dialect by between 0.07 and 0.08 ($\hat{\sigma}_{speaker}$ = 0.08, CrI = [0.07, 0.08]), meaning that speakers differ in their VE ratios by between 32% and 37%. To put this value in context and get an impression of the size of variability across speakers, this value is compared with the degree of variability across dialects. Figure 3.3 illustrates how likely the model deems different degrees of by-speaker and by-dialect variability: highest probability (darker shading) lies where by-dialect variability is greater than by-speaker variability. By the metric of between-dialect variability, Figure 3.3 illustrates that whilst dialects differ in VE size, individual speakers vary little from their dialect-specific baseline value.

3.7 Discussion

The findings from this study will be discussed with respect to the two research questions: (1) how the VE is realised in spontaneous speech, and (2) how the VE varies across dialects and speakers. The VE in English is often considered to be substantially larger than in other languages (Chen, 1970) and claimed to play a significant perceptual role in cueing consonant voicing (Denes, 1955). Taken together, these observations have formed the basis for claims that the VE in English is phonologically specified beyond an otherwise phonetically-consistent acoustic property across languages (Fromkin, 1977; Keating, 1985). Previous work has focused on controlled laboratory speech, leaving open the question of how the VE is realised in spontaneous English speech.

In this study, the overall VE in spontaneous speech was observed to have a maximum size of around 1.2 – substantially smaller than the 1.5 commonly reported in laboratory studies (e.g. House and Fairbanks, 1953; Peterson and Lehiste, 1960; House, 1961; Chen, 1970), and more consistent with previous research on VE in connected speech (Crystal and House, 1982; Tauberer and Evanini, 2009). Spontaneous VE size was also shown to be affected by a range of phonetic factors, such as consonant manner, vowel height, frequency, and speech rate, though the evidence for each of these effects varies substantially (Section 3.6.1). What the effects of these phonetic factors suggest is that contexts where vowels are often shorter also have shorter VE sizes, supporting the argument of 'temporal compression': that vowels which have already undergone shortening cannot be subsequently shortened further (Harris and Umeda, 1974; Klatt, 1976). An interesting exception to this finding is that the VE size was found to be larger for high vowels than non-high vowels in this study (Figure 3.1) – the direction of this effect may be counter to that predicted by temporal compression, and opens a question as to whether this and other predictions of temporal compression are straightforwardly replicable in spontaneous speech environments. The overall smaller-size and impact of phonetic factors of the VE in spontaneous speech indicates a possible fragility of the VE in spontaneous speech, in apparent contrast to the supposed perceptual importance of the VE as a cue to consonant voicing (Denes, 1955; Lisker, 1957; Raphael, 1972). This apparent conflict between the perceptual importance of the VE and its subtlety in production provides an interesting area for future work.

The fact that VE size in English differs so widely between laboratory and connected speech not only demonstrates the importance of speech style and context on phonetic realisation (Labov, 1972; Lindblom, 1990), but also raises the question of 'how big' the VE in English really is, or could be. If larger overall VE size is only observable in laboratory speech, it would be interesting to empirically re-evaulate the question of whether English VE is in fact larger than in other languages. For languages that exhibit smaller VEs than English in laboratory speech (Chen, 1970), it is not clear how such languages may realise the VE in more naturalistic speech. One possibility is that the VE across languages is comparatively small in spontaneous speech and similarly affected by phonetic factors; alternatively, the VE in spontaneous speech across other languages may still be smaller than in English and retain cross-linguistic differences akin to those reported by Chen (1970), and thus English would still retain its status as a language with a distinct realisation of the VE.

The first research question (Section 3.6.1) considered how the VE was modulated in spontaneous speech, averaging across dialects. To what extent dialects themselves differ in VE was the focus of the second research question. As shown in Section 3.6.2, English was shown to exhibit a range of different VE sizes across individual dialects. The dialects with the smallest and largest VEs – Scottish Englishes and AAE, respectively – were expected to show these values given evidence of additional phonological rules governing vowel duration in these varieties (Aitken, 1981; Rathcke and Stuart-Smith, 2016; Holt et al., 2016; Farrington, 2018). Beyond these varieties, dialects appear to differ gradiently from each other, ranging in VE values from around 1.05 in South West England to 1.24 in the Northern Cities region (Figure 3.2). As opposed there being a single 'English'

VE value, there appears to be a range of VE sizes within the language. Such a finding further complicates the notion that English has a particular and large VE relative to other languages. Imagining these different dialects as 'languages' with minimally different phonological structures, this finding demonstrates that such similar 'languages' can have very different phonetic effects (Keating, 1985). This in turn underlies a more nuanced approach to the question of whether English truly differs from other languages in its VE size: not only may English have varieties with greater or lesser VE sizes, but other languages may also exhibit similar dialectal VE ranges.

Individual speakers are also shown to vary in the realisation of the VE, though the extent of this variability is rather limited when compared to variability across dialects (Figure 3.3): that is, whilst dialects appear to demonstrate a range of possible VE patterns, individual speakers vary little from their dialect-specific baseline values. Such a finding supports an interpretation where the VE has a dialect-specific value which speakers learn as part of becoming a speaker of that speech community. The limited extent of speaker variability could predict that the VE will be stable within individual English dialects, given the key role of synchronic speaker variability as the basis for sound change (Ohala, 1989; Baker et al., 2011). This would need checking on a dialect-by-dialect basis, however, given recent evidence of Glaswegian undergoing weakening in its vowel duration patterns (Rathcke and Stuart-Smith, 2016). It also highlights the need for studies addressing both synchronic and diachronic variability across dialects, which we hope to address in future work. One important caveat to this finding is that it assumes that all the dialects analysed in this study contain only speakers who are speakers of that dialect: if a given dialect had a particularly large degree of by-speaker variability, this could reflect the existence of multiple speakers of different dialects (and thus different VE patterns) within that particular dialect coding. This is unlikely to be a particular problem in this study, however, as a separate model that allows for by-speaker variability to vary on a per-dialect basis showed that no dialect with a sufficiently large number of tokens exhibited overly large by-speaker variability (Section 3.6.2).

By using speech data from multiple sources and multiple dialects, it has been possible to investigate variability of a phonological feature across 'English' overall, examine variability at the level of individual dialects and speakers, and reveal the extent of English-wide phonetic variability that was not previously apparent in studies of individual dialects and communities. In this sense, our 'large-scale' approach, using consistent measures and controlling factors, enables us to understand the nature of dialectal variability in the English VE directly within the context of both other dialects and English as a whole.

Whilst this kind of study extends the scope of analysis for (socio)phonetic research, there are of course a number of limitations that should be kept in mind in studies of this kind. This study of the English VE predominantly uses data from automatic acoustic measurements, in turn calculated from forced aligned-segmented datasets. All forced-alignment tools have a minimum time resolution (often 10ms), a minimum segment duration (often 30ms), and there always exists the possibility of poor or inaccurate alignment. This is a necessary consequence of the volume of data used in this study: there is simply *too much* data to manually check and correct all durations, and so the best means of limiting these effects is through sensible filtering and modelling of the data. For example, segments with aligned durations of less than 50ms were excluded, since accurately capturing the duration of a vowel this small could be difficult given the time resolution of the aligner. This decision could exaggerate the size of the VE estimation, as only the most reduced vowels have been removed from the data. Another property of forced alignment which impacts our study of VE is that aligners will only apply the phonological segment label to the segment, meaning that it is possible to only examine VE in terms of *phonological* voicing specification (i.e., whether a segment is underlyingly voiced or not), as opposed to whether the segment itself was realised with phonetic voicing. For example, the realisation of the stop as devoiced (Farrington, 2018) or as a glottal stop (Smith and Holmes-Elliott, 2018), or the relative duration of the closure preceding the vowel (Lehiste, 1970; Port and Dalby, 1982; Coretta, 2019), could affect VE size which is not controllable by exclusively using phonological segment labels. How this kind of phonetic variation, and the more general relationship between a 'phonological' and a 'phonetic' VE, should be understood would certainly be an interesting project for future work. Finally, given the diversity of formats and structures of the corpora available for this study, it has only been possible to categorise and study dialects in a rather broad 'regional' fashion. Similarly, we were unable to investigate the effect of speaker age due to the heterogenous coding of age across the corpora: we agree this is an important dimension that we have attempted to account for in the approach to statistical modelling, and is certainly necessary to examine in future work. Whilst these limitations may be less suitable for approaching other questions in phonetics and sociolinguistics which are concerned with variability at a more detailed level, the approach taken in this study points to a promising first step towards exposing the structures underlying fine-grained phonetic variability at a larger level across multiple speakers and dialects of a language.

3.8 Conclusion

The recent increase in availability of spoken-language corpora, and development of speech and data processing tools have now made it easier to perform phonetic research at a 'large-scale' – incorporating data from multiple different corpora, dialects, and speakers. This study applies this large-scale approach to investigate how the English Voicing Effect (VE) is realised in spontaneous speech, and the extent of its variability across individual dialects and speakers. Little has been known about how the VE varies across dialects bar a handful of studies of specific dialects (Aitken, 1981; Tauberer and Evanini, 2009; Holt et al., 2016). English provides an interesting opportunity to directly examine how phonetic implementation may differ across language varieties with minimally different phonological structures (Keating, 1985). By applying tools for automatic acoustic analysis (McAuliffe et al., 2019) and statistical modelling (Carpenter et al., 2017), it was found that the English VE is substantially smaller in spontaneous speech, as compared with controlled laboratory speech, and is modulated by a range of phonetic factors. English dialects demonstrate a wide degree of variability in VE size beyond that expected from specific dialect patterns such as the SVLR, whilst individual speakers are relatively uniform with respect to their dialect-specific baseline values. In this way, this study provides an example of how large-scale studies can provide new insights into the structure of phonetic variability of English and language more generally.

3.9 Appendices

3.9.1 Dialect-level voicing effect estimates

Table 3.4: Estimated VE sizes (mean, estimated error, and upper & lower credible intervals) for each dialect used in this study.

Dialect	$\hat{\beta}$	Est.Error	95% CrI
Central Scotland	0.99	0.03	[0.93, 1.06]
Edinburgh	1.01	0.04	[0.93, 1.09]
South West England	1.05	0.03	[0.99, 1.12]
Glasgow	1.06	0.02	[1.02, 1.11]
Northern Scotland & Islands	1.06	0.04	[0.99, 1.14]
East England	1.07	0.02	[1.02, 1.12]
Insular Scotland	1.08	0.06	[0.96, 1.21]
Lower North England	1.08	0.03	[1.02, 1.15]
New England	1.08	0.04	[1.00, 1.17]
East Central England	1.09	0.03	[1.03, 1.16]
Scotland	1.10	0.03	[1.04, 1.16]
West Central England	1.11	0.03	[1.04, 1.18]
NYC	1.12	0.04	[1.04, 1.20]
North East England	1.14	0.05	[1.04, 1.26]
Canada (urban)	1.15	0.02	[1.09, 1.21]
Western US	1.15	0.03	[1.09, 1.21]
Canada (rural)	1.17	0.03	[1.12, 1.24]
Ireland	1.17	0.04	[1.07, 1.28]
Philadelphia	1.17	0.02	[1.12, 1.22]
Southern US	1.17	0.03	[1.10, 1.24]
North Midland US	1.18	0.03	[1.11, 1.26]
Northern US	1.18	0.03	[1.11, 1.26]
Wales	1.18	0.03	[1.11, 1.25]
Raleigh US	1.19	0.03	[1.13, 1.26]
South Midland US	1.19	0.03	[1.13, 1.26]
Midwest US	1.20	0.03	[1.14, 1.26]
Northern Cities US	1.24	0.04	[1.15, 1.33]
Rochester NY (AAE)	1.35	0.03	[1.27, 1.44]
Princeville NC (AAE)	1.39	0.03	[1.31, 1.48]
Washington DC (AAE)	1.49	0.02	[1.42, 1.56]

Preface to Chapter 4

Chapter 3 explored how dialects and speakers of English vary in the pre-consonantal voicing effect – the durational difference between vowels preceding voiced and voiceless stops – and observed that this voicing effect was highly variable across English dialects, whilst individual speakers showed a substantially smaller degree of variation. These results are discussed with respect to previous laboratory results, which have largely focused on tightly controlled speech within a single dialect; this chapter has demonstrated that the use of large speech datasets can be useful for 'scaling up' phonetic analyses across multiple related dialects, and was made possible by access to a number of speech corpora and tools to automatically measure acoustic properties of the signal.

As demonstrated by its effectiveness for examining a single phonemic contrast, Chapter 4 applies the same multi-corpus approach from Chapter 3 to explore dialectal variation in another form of phonological contrast in English, and examines how dialects differ systematically in the use of time-dependent acoustic properties in the realisation of particular vowel classes. A substantial literature has examined the role of time-dependent acoustic information in distinguishing vowels within a given linguistic system, whilst studies of vowels across dialects have largely focused on variation in nuclear quality. This chapter compares how different conceptualisations of time-dependent information – multiple formant measurements, representations of formant trajectory, and duration – separately and jointly characterise cross-dialectal variation in time-dependent properties of English vowels.

Chapter 4

Multidimensional acoustic variation in vowels across English dialects

4.1 Introduction

Alongside single-point measurements of vowels, the importance of *time-dependent* information – such as spectral change and duration – in the realisation of vowels has been recognised since the earliest phonetic studies distinguishing vowels in terms of formant values (e.g. Peterson and Barney, 1952; House, 1961; Gay, 1968), with the former noting that "the complex acoustical patterns [...] are not adequately represented by a single section, but require a more complex portrayal." (p.184). Since this time, researchers have developed a range of techniques for characterising dynamic spectral change in vowels, including describing the presence and direction of formant change (Gay, 1970; Labov et al., 1972, 2006), reporting formants from multiple timepoints in the vowel (Hillenbrand et al., 1995; Thomas, 2001; Clopper et al., 2005), to representations of the trajectories themselves (e.g. Watson and Harrington, 1999; Fox and Jacewicz, 2009; Docherty et al., 2015; Renwick and Stanley, 2020). Whilst these studies provide key representations of how dynamic change plays a role in distinguishing vowels within a particular dialect, it remains unclear exactly what role dynamic representations play in characterising how time-dependent properties of vowels vary across a large number of dialects: studies either focus on dialectal differences between closely-related varieties (e.g. within Southern US English, Risdal and Kohn, 2014; Fridland et al., 2014; Farrington et al., 2018; Renwick and Stanley, 2020), or provide relatively broad characterisations of formant dynamics in exchange for wide coverage of differences in regional vowel systems (Labov et al., 2006). Similarly, our understanding of the dialect-specific role of duration has been largely limited to specific investigations of how Southern US speech differs from other English varieties (Clopper et al., 2005; Jacewicz et al., 2007; Fridland et al., 2014), and together suggest that the situation remains unclear as to how best characterise vowel variability across English dialects across mutliple acoustic dimensions.

This study takes an exploratory approach to addressing these issues, and considers two main research questions: *(RQ1) to what extent do time-dependent representations of vowels (formant trajectories, duration) capture additional information (over static F1/F2 position) in describing dialectal variation in vowels?* and *(RQ2) how do measures of formant position, trajectory shape, and duration define the dimensions of vowel variation across dialects of English?* Concretely, RQ1 is addressed through a dialect classification experiment, where different combinations of formant position, trajectory shape, and duration are compared in their ability to correctly classify the dialect of a given vowel. The approach to RQ2 applies dimensionality reduction to determine how these vowel measurements define the main ways in which dialects systematically differ for each vowel. This study takes a 'large-scale' approach, analysing a large amount of acoustic data collected from speech corpora of a range of English dialects. Scaling up the analysis across multiple dialects is made possible by the development of tools for automatic annotation (e.g. Schiel, 1999; Fromont and Hay, 2012; McAuliffe et al., 2017a), acoustic analysis (Rosenfelder et al., 2014; Mielke et al., 2019), and integrating information across idiosyntactic data formats (McAuliffe et al., 2017b, 2019).

The vowels chosen for this study were the following, represented in terms of lexical sets (Wells, 1982): CHOICE, FACE, FLEECE, MOUTH, and PRICE. These vowels were chosen to provide a spectrum of vowels that may vary dialectally by the presence of a glide (Ladefoged and Maddieson, 1993), reflected in the degree of formant change over their timecourse. With the exception of dialects participating in the traditional Southern vowel shift (Labov, 1991; Fridland, 2000; Thomas, 2001; Labov et al., 2006), CHOICE is expected have a diphthongal quality across most dialects. Similarly, whilst FLEECE is known to exhibit some degree of dynamic change (Fox and Jacewicz, 2009; Jacewicz and Fox, 2013; Farrington et al., 2018), it is expected that FLEECE will show relatively little formant change. Diphthongs such as FACE, MOUTH, and PRICE would be expected to vary across dialects in both the degree of dynamic change and overall position, given the presence of both potential 'monopthongisation' of PRICE in Southern US varieties (Thomas, 2001; Labov et al., 2006) and 'Canadian raising' patterns in some Canadian and US varieties within certain phonological contexts (Chambers, 1973; Labov et al., 2006; Boberg, 2008, 2010).

Section 4.2 reviews the literature on the role of vowel dynamics across dialects (4.2.1) and approaches taken to measure dynamic formant trajectories (4.2.2). Section 4.3 discusses the methods used in this study, including the collection and acoustic analysis of the data (4.3.1) and the measures used in the analysis (4.3.2). Section 4.4 reports the results of this study, focusing on the dialect classification experiment (4.4.1) and dimensionality reduction analysis (4.4.2). Section 4.5 provides a discussion of these results in context of previous findings, and Section 4.6 provides a high-level conclusion to the study.

4.2 Background

4.2.1 Dynamic variation in vowels across English dialects

Vowels have long been known to systematically change over their timecourse: even 'nominal' monophthongs – vowels which are considered to be monophthongs in many dialects, such as the vowel in FLEECE – move in formant space to a position distinct from their onsets (Joos, 1948; Assmann et al., 1982; Fox, 1983; Nearey and Assmann, 1986; Hillenbrand et al., 1995). Spectral change in vowels also plays a role in the development of dialect-specific shifts: for example, the weakening of the glide in PRICE vowels, long been considered a hallmark of speech in Southern US varieties since the early twentieth century (e.g. Evans, 1935), is considered to be the starting point of the 'Southern Vowel Shift' (SVS), in which the front diphthongs (FACE, FLEECE) fall and front lax vowels (DRESS, KIT) develop diphthong-like status (Labov, 1991; Clopper et al., 2005; Labov et al., 2006). Thomas (2001, 2003) notes that, far from there being a single glide weakening pattern, different regional communities in the South exhibit distinct gradient patterns of glide weakening, underlining the importance of detailed phonetic information about glide dynamics beyond a binary marker of glide presence (Farrington et al., 2018). Similarly, the weakening of the glide in MOUTH vowels is widespread in Western Pennsylvania, Cockney, and South African Englishes (Johnstone et al., 2002, 2015).

Cross-dialectal studies of spectral (formant) change have observed some degree of differences between varieties: Clopper et al. (2005) observe that Southern US speakers produce a greater degree of spectral change in contrasting tense and lax vowels, whilst Midland US speakers distinguish KIT and DRESS predominantly through spectral change than in spectral position. Fox and Jacewicz (2009) observe greater spectral change from Southern US speakers in the production of KIT vowels compared with Ohio speakers, whilst North Carolina speakers engage in far less spectral change in the realisation of PRICE, reflecting the presence of glide weakening in Southern US regions (see also Jacewicz and Fox, 2013). Farrington et al. (2018) examine how three Southern US varieties (Tennesee, North Carolina, Virginia) participate in SVS-like patterning through the application of dynamic measures, and observe that dialects substantially differ in the dynamic realisation of the front vowel system. Williams et al. (2019) observe that differences between British and Australian English varieties are better predicted by explicit measures of the formant trajectory than the average of all points in a trajectory, whilst both Risdal and Kohn (2014) and Swan (2016) report subtle distinctions between dialects of African American English and Canadian English in dynamic formant shape that would have been otherwise obscured through exclusive analysis of formant position.

Durational differences *between* dialects also remain relatively understudied, though there is a substantial literature that has been focused on durational differences between the vowels of English within a given dialect (e.g. House, 1961; Klatt, 1973; Umeda, 1975; Crystal and House, 1982). The vast majority of the cross-dialectal literature on duration has been concerned with whether Southern US speakers have overall longer vowel durations than other regions (i.e., the 'Southern drawl', Bailey, 1968; Wetzell, 2000). Clopper et al. (2005) find that the dialectal difference in duration is retained exclusively for lax vowels, linked to their more peripheral quality as part of the SVS (Labov et al., 2006). Similarly, Southern US vowels were observed to be longest in a study comparing the South, the Inland North, and Midland speakers (Jacewicz et al., 2007), which is also linked to increased degree of spectral change (Fridland et al., 2014). In a study of connected interview speech, Tauberer and Evanini (2009) find that Southern US speakers produce longer vowels than Northern speakers in spite of having similar speech rates (Clopper and Smiljanic, 2011).

4.2.2 Approaches to the measurement of formant dynamics

In attempting to capture dialectal variation in spectral change, researchers have applied a number of different techniques. Studies of monophthongs attempt to capture the nuclear quality of the vowel by measuring some steady central point, reflecting the point where the formants are least likely to be affected by coarticuation from surrounding segments (Labov et al., 2006; Thomas, 2018). For diphthongs, this could involve additional points of formant measurement in the midpoint or vowel offset (Hillenbrand et al., 1995; Thomas, 2001; Clopper et al., 2005) or noting the presence and direction of the glide auditorily (Labov et al., 1972, 2006).

Alongside this approach of reporting formants at two to three timepoints, a range of approaches have been applied that directly reference the shape of the formant trajectory itself. One approach involves fitting a parametric curve to the trajectory using Discrete Cosine Transform (DCT). DCTs are convenient for modelling a trajectory as the model coefficients represent increasingly complex representations of the trajectory shape: the zeroth coefficient captures the mean value of all points; the first coefficient provides the direction and magnitude of the slope, and the second coefficient represents the degree of curvature in the trajectory (Morrison, 2013). The zeroth and first coefficient have been found to be most informative in classifying vowels within a given system, whilst the inclusion of higher-order representations in the second coefficient only provide marginal increase (Zahorian and Jagharghi, 1993; Watson and Harrington, 1999; Hillenbrand et al., 2001; Williams and Escudero, 2014). Other curve-fitting approaches, such as Smoothed-Splines ANOVA (Docherty et al., 2015), Functional Data Analysis (Risdal and Kohn, 2014; Gubian et al., 2015), and generalised additive mixed models (GAMMs), have also been applied to the analysis of formant shape variation (Cole and Strycharczuk, 2019; Kirkham et al., 2019; Renwick and Stanley, 2020). As GAMMs are best used in making comparisons between two groups (Sóskuthy, 2017), they do not straightforwardly lend themselves to performing simultaneous comparisons across a large number of dialects.

Another commonly-used set of measures derives from calculations of 'vowel section length' (VSL): the Euclidean distance between two formant points (n, m):

$$VSL_{n,m} = \sqrt{(F1_n - F1_m)^2 + (F2_n - F2_m)^2} \tag{4.1}$$

A measure of the overall spectral change (called 'Vector Length') is derived from calculating the VSL of the vowel onset and offset, whilst more complex representations of the trajectory can be derived from the summation of VSLs calculated from subsets of the points, such as onset to midpoint + midpoint to offset (Fox and Jacewicz, 2009). This approach has been most-applied in studies of dialectal variation in English (Fox and Jacewicz, 2009; Cardoso, 2015; Farrington et al., 2018) and other languages (Mayr and Davies, 2011; Schoorman et al., 2015). Whilst these methods have not been explicitly compared, the decision to make sure of the vector-based measurements in this study is based around the relative comparaibility with the previous cross-dialectal work using this measure, as well as its relative interpretability as a representation of spectral change.

4.3 Methods

4.3.1 Data

The data used in this study was collected as part of the SPeech Across Dialects of English (SPADE) project (Sonderegger et al., 2020b, https://spade.glasgow.ac.uk/), which aims to analyse and explore phonological and phonetic variation over time and geographic location in British and North American English varieties (e.g. Mielke et al., 2019; Stuart-Smith et al., 2019; Tanner et al., 2019b). This is enabled by the collection of a wide array of speech corpora from diverse sources, and the integration of multiple corpora into a cross-dialectal analysis. In this study, North American dialects refers to dialects in the countries of Canada and the United States as outlined in *The Atlas of North American English* (Labov et al., 2006), based on phonological isoglosses such as the participation in regional mergers or chain shifts. Due to the relative sparsity of Canadian data available compared with United States and British dialects, Canadian dialects in this study were distinguished along rural and urban lines instead of geographical location (Greenbaum and Nelson, 1996; Rosen and Skriver, 2015). Dialectal distinctions for British English dialects were based on Trudgill's (1999) modern dialectal groupings, based on both phonological and lexical distinctions. For Scottish dialects specifically, speakers were grouped based on information from *The Scottish National Dictionary*.[1] The data in this study was extracted from a total of 11 corpora, which are outlined below.

- *Audio British National Corpus* (Audio BNC, Coleman et al., 2012): The spoken section of the British National Corpus, a corpus of over 1000 speakers engaging in spontaneous speech in informal (radio shows, phone calls) and formal (business & government) speech. Due to a range of issues concerning the audio quality of the Audio BNC (overlapping speech, background noise, artefacts from microphone handling, etc.), substantial portions of the corpus are not accurately aligned. In order to derive a subset of the Audio BNC that was accurately aligned, inclusion criteria for aligned utterances was applied. These criteria were the following: the utterance must be longer than one second, the utterance must contain at least two words, the mean harmonics-to-noise ratio of the recording was at least 5.6, and the mean difference in segmental boundaries between the original alignment and a subset re-aligned with the Montreal Forced Aligner (MFA, McAuliffe

[1] Part of *The Dictionary of the Scots Language*, https://dsl.ac.uk/).

et al., 2017a) was no greater than 30ms.[2] As a manual check, 50 Praat TextGrids containing 'acceptable' alignment were manually checked by the first author and deemed equivalently accurate to other forced alignment techniques.

- *Buckeye* (Pitt et al., 2007): Spontaneous interview speech with 40 speakers from the Columbus, Ohio region, recorded in 1990s-2000s. The corpus contains hand-corrected segmental boundaries with phonetic transcription labels. In order to make the transcription of Buckeye corpus comparable with the other corpora (which are phonologically transcribed), the phonetic labels were converted back to their underlying (phonological) transcriptions provided along with the corpus.

- *Canadian Prairies* (Rosen and Skriver, 2015): Spontaneous sociolinguistic interviews with speakers of various ethnic backgrounds in the Canadian provinces of Alberta and Manitoba, recorded between 2010 and 2016. Dialects were split into 'urban' if dialects were classed as living in an urban or semi-urban environment, and 'rural' otherwise. This corpus was aligned with the MFA.

- *Hastings* (Holmes-Elliott, 2015): Spontaneous sociolinguistic interviews with 46 speakers from Hastings (South East England), aligned with the FAVE aligner (Rosenfelder et al., 2014).

- *International Corpus of English – Canada* (ICE-Canada, Greenbaum and Nelson, 1996): Interview and broadcast speech, recorded in the 1990s, with speakers from various regions in Canada, aligned with the MFA. For this corpus, speaker dialect was defined as 'urban' if the speaker metadata referred to the speaker's birthplace as a major Canadian city (e.g., Montreal, Toronto, Ottawa, Vancouver, etc), and 'rural' otherwise.

[2] We are grateful to Michael Goodale for devising and implementing this inclusion protocol.

- *Intonational Variation in English* (IViE, Grabe, 2004): Recordings of English, Welsh, Scottish, and Irish speakers engaging in structured spontaneous and reading tasks, recorded 1998-2000. Aligned using the MAUS aligner (Schiel, 1999). This study used the English read-speech subset of the data, using speakers from London, Cambridge, Newcastle, Bradford, and Leeds.

- *Modern RP* (Fabricius, 2000): Read speech by Cambridge University students in the 1990s. Speakers were defined as having 'upper middle-class' accents and were selected based on whether one of their parents worked in a professional occupation and the speaker had attended private school. The corpus was aligned with FAVE.

- *Raleigh* (Dodsworth and Kohn, 2012): Semi-structured sociolinguistic interviews with 59 white speakers from Raleigh, North Carolina, born between 1955 and 1989. Aligned with the MFA.

- *The Scottish Corpus of Texts and Speech* (SCOTS, Anderson et al., 2007): Approximately 260 spoken texts, ranging from spontaneous informal conversation to structured interviews, recorded during the 2000s and aligned with the MFA.

- *The Sounds of the City* (SOTC, Stuart-Smith et al., 2017): Vernacular and standard Glaswegian English, recorded 1970s-2000s, of 142 speakers. Sourced from historical archives and sociolinguistic interviews, and aligned with LaBB-CAT (Fromont and Hay, 2012).

- *Switchboard* (Godfrey et al., 1992): 2400 spontaneous telephone conversations between unfamiliar participants from multiple dialect regions in the United States, consisting of approximately 500 speakers.

Vowel tokens were extracted from each of the corpora using the ISCAN software (McAuliffe et al., 2019), which processes audio and textgrid files (of various

Table 4.1: Speaker and token count for each dialect used in this study, separated by the corpus from which the data was originally sourced.

Continent	Dialect	Corpus	Speakers	Tokens
North America	Canada (rural)	Canadian-Prairies	44	20042
	Canada (rural)	ICE-Canada	8	2764
	Canada (urban)	Canadian-Prairies	67	38021
	Canada (urban)	ICE-Canada	8	877
	Midwest US	Buckeye	40	17669
	New England	Switchboard	18	2868
	North Midland US	Switchboard	44	7126
	Northern US	Switchboard	53	7494
	NYC	Switchboard	19	3183
	Raleigh US	Raleigh	100	64659
	South Midland US	Switchboard	106	20327
	Southern US	Switchboard	37	5595
	Western US	Switchboard	45	6376
United Kingdom	Central Scotland	SCOTS	23	5237
	East Central England	Audio BNC	30	3877
	East England	Audio BNC	100	13429
	East England	Hastings	49	25477
	East England	IViE	12	972
	East England	IViE	11	992
	East England	ModernRP	48	2811
	Edinburgh	SCOTS	18	2361
	Glasgow	SCOTS	26	4432
	Glasgow	SOTC	155	45487
	Lower North England	Audio BNC	41	5445
	Lower North England	IViE	11	891
	Lower North England	IViE	10	760
	North East England	Audio BNC	10	917
	North East England	IViE	12	1018
	Northern Scotland & Islands	SCOTS	31	3998
	South West England	Audio BNC	37	3458
	West Central England	Audio BNC	32	4497
Total	**21**	**11**	**1245**	**323060**

forced alignment formats) and returns CSV files based on user-defined filtering criteria. In this study, only primary-stressed vowels were included, and only vowels of the five types (CHOICE, FACE, FLEECE, MOUTH, PRICE), defined using the UNISYN cross-dialect lexicon (Fitt, 2001), were analysed. The set of tokens consisting of the PRICE vowel does not include the PRIZE subset, which would be expected to behave as a separate class in both Scottish, Canadian, and some US dialects (Labov, 1963; Chambers, 1973; Aitken, 1981). Tokens with a duration shorter than 50 milliseconds were not extracted, in line with previous studies of vowel formants (Dodsworth, 2013; Fruehwald, 2013). Vowels with a duration longer than 500 milliseconds were excluded from further analysis.

Formants were extracted in Hertz at 21 equally-spaced points, and were automatically measured with PolyglotDB (McAuliffe et al., 2017b) using the mea-

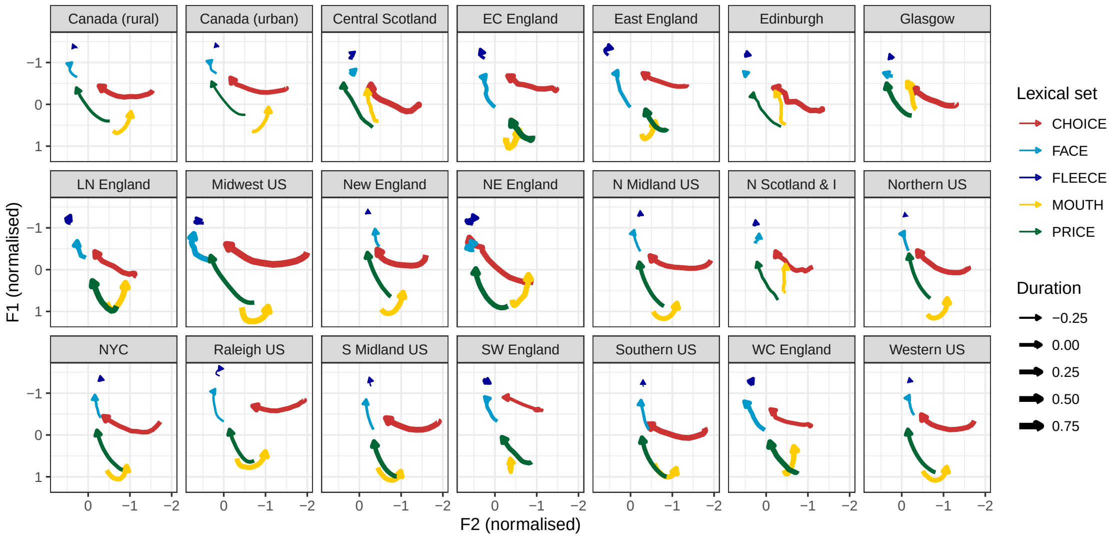

Figure 4.1: Normalised by-dialect vowel trajectories for the central 60% of the five vowels analysed, averaged over all tokens for that dialect. Duration corresponds to the within-speaker Z-score normalisation.

surement scheme described in Mielke et al. (2019), which is based on the measurement system in FAVE-Extract (Rosenfelder et al., 2014). Each formant was measured multiple times with varying numbers of LPC coefficients; each candidate measurement was compared with a corpus-specific prototype for that vowel, consisting of mean F1-3 formants, B1-3 bandwidths, and covariance matrices, based on a point taken from 33% within the vowel. The candidate with the smallest Mahalanobis distance from the prototype was selected as the formant measurement. The corpus-specific prototypes were generated by first measuring formant values based on a prototype which shared the same phone label set. Raleigh (Dodsworth and Kohn, 2012) was used for North American dialects using CMUDict labels, and Santa Barbara (Bois et al., 2000) for North American dialects with SAMPA labelling. For British English dialects, SOTC (Stuart-Smith et al., 2017) and Modern RP (Fabricius, 2000) were used for Scottish and English CMUDict-labelled corpora respectively, with the Edinburgh 'Arthur the Rat' corpus was used as a prototype for all British English SAMPA corpora. The first and last 20% of the vowel was excluded to minimise the influence of surround-

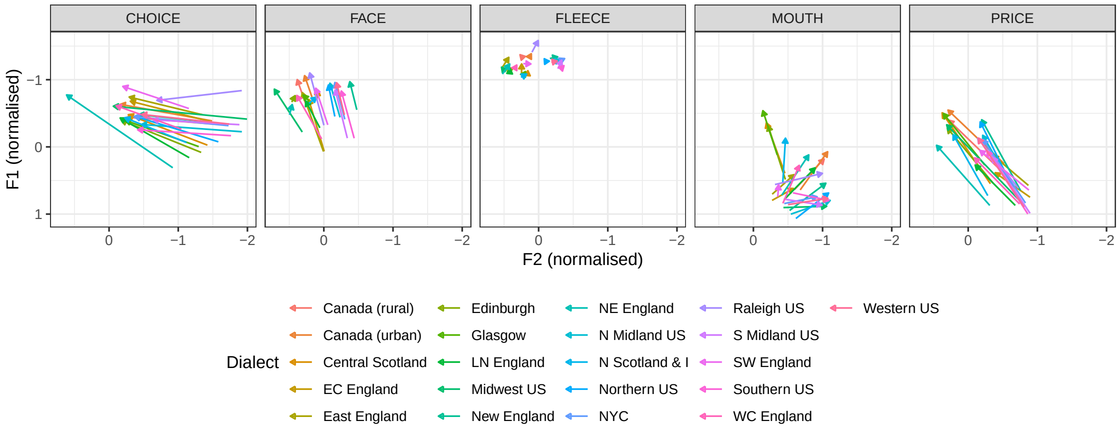

Figure 4.2: Mean dialect F1 and F2 values for the 5 vowels (CHOICE, FACE, FLEECE, MOUTH, PRICE). One point per dialect. Onset value represented by the start point of the arrow; offset represented by position of arrowhead.

ing segments (Fox and Jacewicz, 2009; Williams and Escudero, 2014; Risdal and Kohn, 2014; Williams et al., 2019). The remaining middle 60% of the vowel (13 points) was then Z-score normalised against all vowels produced by the speaker, including those not analysed in this study ('Lobanov normalisation', Lobanov, 1971). In total, 323060 tokens (6259 types), corresponding to 1245 speakers from 21 dialects of North American and British English, were used in this study (Table 4.1). Figure 4.1 illustrates the averaged vowel trajectories and duration for each vowel within each dialect.

4.3.2 Measures

The goal of this study is to examine how measures of formant position, trajectory shape, and duration together inform the dimensions of dialectal variation in English vowels. We derived 8 measures, with 4 measures capturing the position of the vowel's position in formant space, 3 measures characterising the shape of the trajectory independent of formant position, and 1 measure of vowel duration.

Formant position: F1, F2 onset & offset

To capture the formant position, the speaker-normalised F1 and F2 values were taken from the 20% and 80% points, corresponding to the vowel **Onset** and **Offset** respectively. Figure 4.2 illustrates the position of the onset and offset of each dialect, for each of the five vowels included in this study. As can be seen in this figure, dialects appear to differ substantially in their overall position in formant space; the degree of this difference, however, varies across each vowel. For example, dialects are somewhat diffused for CHOICE (outer left) FACE, (inner left), and PRICE (outer right), whilst maintaining some similarity in the difference between the onset and offset (reflected in the direction of the arrow) across dialects. One set of exceptions to this, however, are a selection of Scottish and North English dialects, demonstrating relatively static realisations of FACE (Haddican et al., 2013). For FLEECE (centre), most dialects vary in their formant position, but show little formant change over the timecourse of the vowel. For MOUTH, in contrast, a number of Scottish dialects (e.g., Central Scotland, Edinburgh, Glasgow) show distinct differences in both their slightly higher starting position and a distinct upwards movement; this reflects the distinct front-raising realisation of MOUTH in Scottish English dialects (Stuart-Smith, 2008).

Trajectory shape: Vector Length, offset, & angle

Three measures were calculated to capture properties of a vowel's formant trajectory independent of its position in formant space. The first, **Vector Length** (calculated from VSL, Equation 4.1), was calculated between the onset and offset value, reflecting the overall degree of linear spectral change over the vowel's timecourse. One measurement commonly used in studies of trajectory shape, trajectory length (Fox and Jacewicz, 2009; Mayr and Davies, 2011; Schoorman et al., 2015; Farrington et al., 2018; Holt and Ellis, 2018) is calculated as the summation of two VSLs: one measuring the distance from the vowel onset to mid-

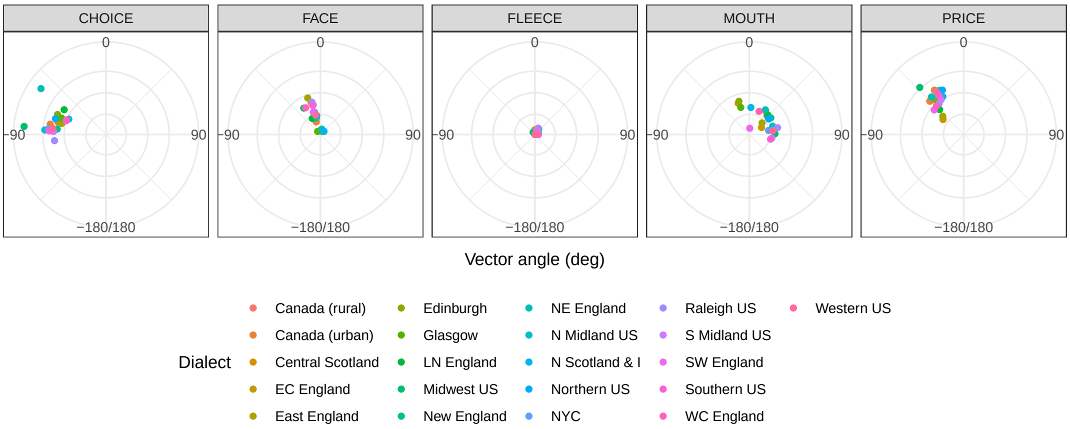

Figure 4.3: Mean dialect values for Vector Angle (direction on compass) and Vector Length (distance from centre), for each of the five vowels in the study. One point per dialect.

point, and another measuring the distance from the midpoint to the vowel offset. As trajectory length is highly correlated with Vector Length ($r = 0.99$, $p < 0.001$ for this data), we derived **Vector Offset**, which is trajectory length subtracted from Vector Length, reflecting the residual difference between the two measures. Finally, **Vector Angle**, a measure of a vowel's *direction* of change, was derived by calculating the arctangent of the onset and offset position, converted to degrees, and then placed on a $180/-180^{\circ}$ scale by adding 360 to Vector Angles with values less than 0° and subtracting 360 from Vector Angles with values greater than 180°.

Figure 4.3 illustrates the dialectal variation in both Vector Length (a dialect's distance from the centre of the compass) and Vector Angle (the orientation around the compass). This figure demonstrates that, as with formant position (Fig. 4.2), the degree of dialectal variation for these dimensions differs by individual vowel. Specifically, vowels such as CHOICE (outer left) and MOUTH (inner right) show wide dialectal variation in Vector Angle, exemplified by the spread of dialects around the $180/-180^{\circ}$ compass. In contrast, FACE and PRICE vowels show little dialectal variation in Vector Angle; instead, dialects appear to vary in their overall degree of spectral change. FLEECE shows very little overall spectral change,

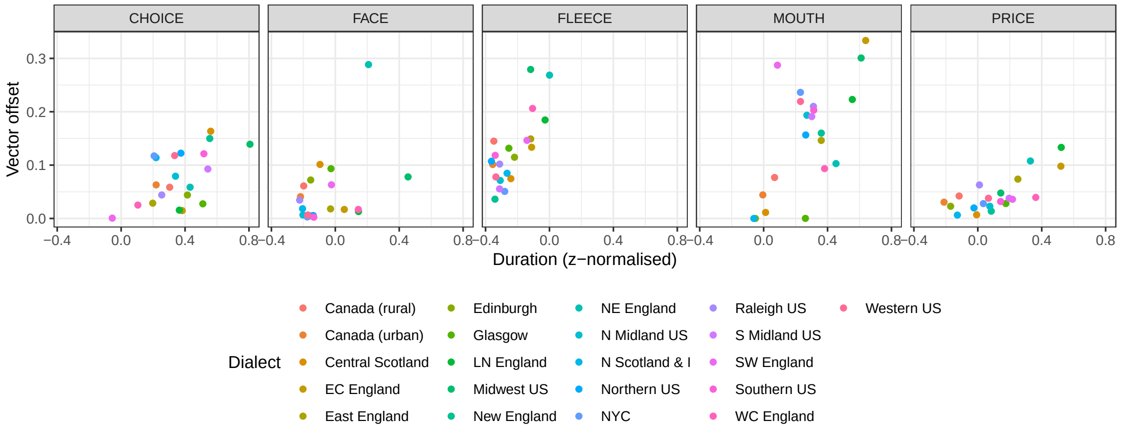

Figure 4.4: Mean dialect values for z-normalised vowel duration (x-axis) and Vector Offset (y-axis), for each vowel (CHOICE, FACE, FLEECE, MOUTH, PRICE). One point per dialect.

reflected in all dialects clustered around the centre of the compass. Figure 4.4 (y-axis) shows the dialectal variation in Vector Offset, and illustrates the wide difference in dialectal variation across vowels. For example, PRICE (outer left) shows relatively low values for Vector Offset, as well as little dialect variation, suggesting that spectral change in PRICE vowels is relatively linear. In contrast, FLEECE and MOUTH show much greater and more variable Vector Offset patterns; since the overall degree of linear spectral change is low for FLEECE (Fig. 4.3, centre), this would suggest that FLEECE vowels may undergo substantial *non-linear* change, returning to a position in formant space similar to its onset position. Similarly, MOUTH shows a wide range of dialectal variability, from Scottish and Canadian dialects demonstrating minimal non-linear change, up to dialects such as East England, Lower North England, and West Central England exhibiting substantial non-linear spectral change.

Duration

Vowel duration was calculated by Z-score normalising the vowel's force-aligned duration against all of the speaker's vowels (including vowels not analysed in this study). Figure 4.4 (x-axis) shows the dialectal distribution of vowel durations for each of the five vowels. As with previous measures, duration exhibits a wide range of variability across dialects, but this variability is somewhat structured by individual vowels. PRICE vowels (outer right), for example, demonstrates substantial variation in duration across dialects, with Glasgow and East Central England returning the longest average duration across sampled dialects. In contrast, FLEECE (centre) shows a much tighter distribution in duration values across dialects.

4.4 Results

The goal of this study is to address the role of formant position, trajectory shape, and duration in describing how vowels vary across dialects of English. Specifically, the first research question (RQ1) considers the extent to which time-dependent information about trajectory shape and duration provide additional information about dialectal variability on top of measures of formant position, and is addressed through a classification experiment (Section 4.4.1), where different combinations of measures are used to train a supervised learning model to predict the dialect label associated with speakers. The second research question (RQ2) aims to address how these measures may combine to represent underlying patterns of variation and represent the primary dimensions of dialectal variability (Section 4.4.2), and is performed through the use of dimensionality reduction.

4.4.1 Dialect classification experiment

To address RQ1, quantifying the relative roles of static and dynamic measurements of vowels across dialects, support vector machines (SVMs) were trained on each vowel using the `e1071` package (Meyer et al., 2019) in *R* (R Core Team, 2019). SVMs are a class of supervised learning model, which can be trained to assign ('classify') a label to an example, given values of a predictor, such as providing a dialect label (e.g., Southern US, Glasgow) to a datapoint based on provided formant, trajectory, and duration values. This is achieved through attempting to separate points along a decision boundary: this decision boundary can be either linear, or non-linear by mapping the decision boundary into higher-dimension space through using different kernel functions (Kuhn and Johnson, 2013). The radial basis function (RBF) kernel was used for SVMs in this study, which allows for fitting non-linear decision boundaries, using parameters set at their default values in the `svm` function, with the exception of C and γ which were tuned (see below). The use of a non-linear boundary is useful for this experiment, as the acoustic dimensions being used as input do not distribute positive ('dialect') and negative ('not dialect') cases as either side of a hyperplane without a non-linear mapping. As SVMs can be used for multiclass classification – providing a label from a set of > 2 possible labels – they are well suited to addressing RQ1: specifically, an SVM can be trained to predict one of N-many possible dialect labels given prototypical formant position, trajectory shape, and duration values.[3]

The data was prepared for SVM training by averaging formant, trajectory shape, and duration values for each speaker across each vowel, and separate SVMs were trained for each of the 5 vowels analysed in this study. The choice to use one observation per speaker (compared to one value for each observation in the dataset) was motivated by the desire to abstract away from variability due to phonological environment, and instead achieve an 'average' value for a vowel for that speaker by averaging over all observations of that vowel by that speaker. To examine how different combinations of measures best contribute to accurately predicting the dialect, 5 SVMs (one for each vowel) were trained on a different set of measurements:

[3] Whilst SVMs were chosen for this study, other classification methods would also have been suitable for performing this experiment. Linear discriminant analysis may not have been appropriate given the assumptions required for the method (such as assuming equal variance for each of the dialect labels, assuming a linear boundary between labels), but other methods such as random forests or k-means clustering would serve as alternatives to SVMs.

1. Formant values (F1/F2 onset + offset)
2. Trajectory shape (Vector Length, offset, angle)
3. Duration
4. Formants + duration
5. Trajectory + duration
6. Formants + trajectory
7. Formants + trajectory + duration

Each SVM was trained on a 80% subset of the data, and tuned to derive the best parameters (margin parameter C, kernel parameter γ) by 10-fold cross validation using the `tune` function.[4] As the dialect groups are unbalanced in this data (some dialects have more speakers than others: Table 4.1), the performance on the 20% test set is evaluated using a metric that appropriately accounts for class imbalance. This measure, balanced accuracy, accounts for class imbalance by normalising the true positive and negative rates by the relative number of samples (Kelleher et al., 2015). Specifically, this is calculated as the average of sensitivity (portion of correctly-predicted 'positive' values) and specificity (portion of correctly-predicted 'negative' values), averaged using 'weighted macro-averaging'. This calculates the overall accuracy as the sum of binary 'one versus all' accuracy for all dialect labels, where a given speaker is classified as either belonging to the dialect under consideration (positive) or not (negative), and then weighted by the overall number of samples in each class. Balanced accuracy was calculated using the `yardstick` package (Kuhn and Vaughan, 2020). To directly compare how different combinations of metrics aid in the classification of dialects, the differences in balanced accuracy for each vowel was calculated,

[4] The ranges provided for C and γ were both set as $\{10^{-10}, 10^{-9}, 10^{-8}, ..., 10^{10}\}$. Best values for C, γ parameters for each SVM can be found in Appendix 4.7.1.

Table 4.2: Balanced accuracy (Bacc., %) for each SVM, trained with different configurations of formant position, trajectory shape, and duration measures.

Measures	**CHOICE**	**FACE**	**FLEECE**	**MOUTH**	**PRICE**
Formants (F1, F2 onset + offset)	57	61	58	61	62
Trajectory (Vector Length, offset, angle)	55	62	55	64	56
Duration	52	53	55	53	57
Formants + duration	60	65	62	66	66
Trajectory + duration	56	65	57	65	61
Formants + trajectory	58	63	61	67	65
Formants + trajectory + duration	58	64	63	70	69

Table 4.3: Differences in balanced accuracy (ΔBacc.) between different combinations of measurements, with within-vowel FDR-adjusted p-values calculated using a one-sided permutation test with 1000 permutations.

	CHOICE		**FACE**		**FLEECE**		**MOUTH**		**PRICE**	
Comparisons	**ΔBacc.**	p	**ΔBacc.**	p	**ΔBacc.**	p	**ΔBacc.**	p	**ΔBacc.**	p
Formants − Trajectory	0.02	0.12	−0.01	N/A	0.03	< 0.05	−0.02	N/A	0.06	< 0.001
{Formants, Duration} − Duration	0.08	< 0.001	0.12	< 0.001	0.07	< 0.001	0.14	< 0.001	0.09	< 0.001
{Formants, Trajectory} − Formants	0.01	0.20	0.01	0.19	0.03	< 0.05	0.06	< 0.001	0.03	< 0.05
{Formants, Trajectory} − Trajectory	0.03	0.08	0.01	0.29	0.06	< 0.001	0.04	< 0.05	0.09	< 0.001
{Formants, Trajectory, Duration} − {Formants, Duration}	−0.01	N/A	−0.01	N/A	0.01	0.13	0.04	< 0.05	0.03	< 0.05
{Formants, Trajectory, Duration} − {Trajectory, Duration}	0.03	0.06	−0.01	N/A	0.07	< 0.001	0.05	< 0.001	0.09	< 0.001
{Formants, Trajectory, Duration} − {Formants, Trajectory}	0.19	0.13	0.01	0.19	0.03	< 0.05	0.03	< 0.05	0.04	< 0.001

and significance of the difference was evaluated through a one-sided permutation test, comparing the likelihood of whether the difference was greater than the average difference observed for 1000 permutations (Table 4.3), and were subject to within-vowel Benjamini-Hochberg False Discovery Rate (FDR) adjustment for multiple comparisons.

Table 4.2 (visualised in Fig. 4.5) shows the classification performance, evaluated with balanced accuracy, for each SVM. These results show that all measures return balanced classification accuracy just over chance: one possibility is that classification of dialect labels is a difficult task due to substantial overlap of values for each dialect (for a similar interpretation, see Williams et al., 2019), or that these measures are simply not the optimal ones for discriminating dialects. Despite this, some observations can be made about the relative performance of single, and combinations, of acoustic measures, in classifying dialect variation.

Considering the use of each type of measure in isolation (Table 4.2, rows 1-3), formant position provides a modest but consistent improvement in performance

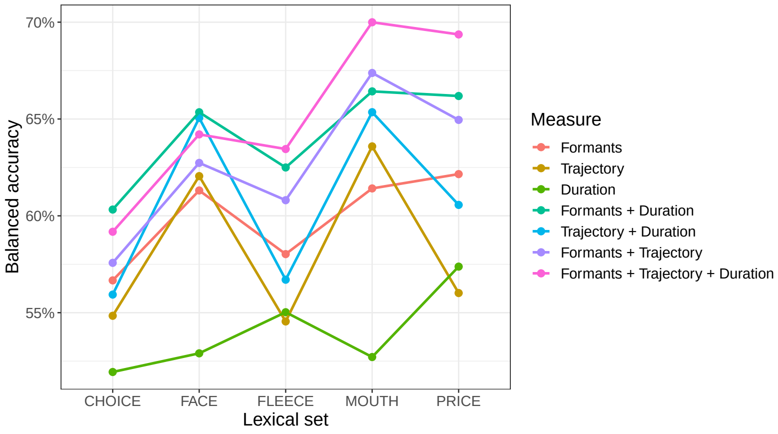

Figure 4.5: Balanced accuracy values for each combination and vowel measurement (as reported in Table 4.2).

over measures representing the shape of the formant trajectory (Table 4.3, row 1). Duration returns the lowest overall performance, suggesting that duration itself is not the primary means by which dialects are distinguished for these vowels. Duration does, however, provide significant additional accuracy when included alongside either formant position of trajectory (Table 4.2, rows 4-5), improving over their respective performances without duration (Table 4.3, row 2).

Using both formant position and trajectory shape to predict dialect labels (Table 4.2, row 6) also significantly improves over the use of each measure in isolation: this improvement is greater and more consistent when compared to trajectory information in isolation (Table 4.3, row 4), suggesting that trajectory information provides additional information over formants (cf. Fox and Jacewicz, 2009; Farrington et al., 2018; Williams et al., 2019), but crucially that the inclusion of formant position improves classification accuracy over trajectory information. The combination of all measures – formant position, trajectory shape, and duration – generally returns the highest performance (Table 4.2, row 7). Comparing these with any of the subset combinations (Table 4.3, rows 5-7), adding trajectory information to formants and duration only provides a modest improvement, whilst adding formant values to predictions using trajectory and duration information results in a greater performance increase.

Table 4.4: Loadings for the first (PC1) and second (PC2) principal components, standard deviation, and variance (proportional, cumulative) for each vowel, with three largest loadings in bold font.

	CHOICE		FACE		FLEECE		MOUTH		PRICE	
Measures	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**
F1 onset	0.31	**−0.46**	0.20	**−0.54**	0.33	**0.47**	0.40	0.16	0.28	**0.46**
F1 offset	−0.24	−0.31	−0.31	−0.23	0.25	**0.67**	**0.46**	0.02	**0.47**	0.02
F2 onset	**0.52**	0.05	**−0.51**	−0.03	**0.44**	−0.18	−0.04	**−0.68**	−0.34	−0.36
F2 offset	**0.44**	−0.26	**−0.41**	−0.28	**0.45**	−0.16	−0.37	**−0.41**	**−0.36**	**−0.38**
Vector length	−0.06	**−0.43**	0.34	**−0.45**	0.04	**−0.45**	−0.24	**0.42**	−0.33	0.29
Vector angle	**0.50**	−0.17	−0.21	0.36	−0.26	0.20	**0.42**	0.13	−0.23	**0.54**
Vector offset	−0.30	−0.37	**−0.45**	0.00	0.41	−0.18	**0.41**	−0.27	0.35	−0.35
Duration	−0.20	**−0.52**	−0.27	**−0.49**	**0.43**	−0.07	0.30	−0.28	**0.42**	−0.13
Std. deviation	1.84	1.55	1.84	1.57	2.04	1.27	2.13	1.38	2.03	1.38
Prop. variance (%)	42.1	30.1	42.1	30.1	52.2	20.2	56.6	23.8	51.6	23.9
Cum. variance (%)		72.2		73		72.4		80.4		75.5

Looking at the average performance across vowels, CHOICE and FLEECE vowels consistently return the lowest performance (Table 4.2, Figure 4.5) and smallest differences between different measurement configurations (Table 4.3). PRICE and MOUTH vowels, in contrast, generally return the highest dialect classifications for all of the vowels, and are more likely to demonstrate differences between each set of measurements.

4.4.2 Principal component analysis

The second research question (RQ2) moves on to consider how measures of formant position, trajectory shape, and duration together capture the key dimensions of variation across English dialects. This question is addressed here by applying dimensionality reduction: a technique by which the number of individual dimensions (each measure in this case) is reduced to a smaller set of composite dimensions which represent the principal directions of variation. This study applies Principal Component Analysis (PCA), a popular dimensionality reduction technique, which linearly maps the set of variables into a set of orthogonal (uncorrelated) 'principal components'. The results of PCA are typically analysed in terms of a principal component's *loadings*, which represent coefficients of each

Table 4.5: Contributions of each measure in accounting for the variance in the first two principal components (%).

	CHOICE		**FACE**		**FLEECE**		**MOUTH**		**PRICE**	
Measures	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**
F1 onset	9.5	**21.0**	4.1	**28.7**	11.2	**22.1**	15.7	2.5	7.7	**20.9**
F1 offset	5.5	9.6	9.6	5.5	6.3	**44.5**	**21.0**	0.0	**22.1**	0.0
F2 onset	**27.3**	0.3	**25.7**	0.1	19.7	3.2	0.1	**45.8**	11.7	13.2
F2 offset	**19.0**	7.0	**17.1**	7.8	**20.5**	2.4	13.4	**16.9**	**12.8**	**14.3**
Vector length	0.4	**18.5**	11.5	**20.6**	0.2	**20.0**	6.0	**17.4**	11.2	8.2
Vector angle	**25.2**	3.0	4.3	13.0	6.8	4.0	**17.7**	1.7	5.2	**29.4**
Vector offset	8.9	13.8	**20.5**	0.0	**17.1**	3.3	**17.1**	7.6	12.0	12.2
Duration	4.1	**26.7**	7.2	**24.3**	**18.2**	0.5	9.1	8.1	**17.3**	1.8

measure within the component (Table 4.4), and *contribution*, which captures how much a given measure accounts for the variance within a given component (Table 4.5). Loadings with the same sign are positively correlated within a principal component, whilst those with opposing signs are negatively correlated; a larger loading value for a given measure indicates a greater importance for the measure in the principal component, which is also reflected in the measure's contribution. PCA was performed using the `FactoMineR` package (Lê et al., 2008), and all measures were scaled and centred (necessary as PCA results are sensitive to the scale of predictors).

Loadings for each acoustic dimension for the first (PC1) and second (PC2) principal components for each vowel can be seen in Table 4.4. The relative contribution for each measure to principal components can be seen in Table 4.5, and the position of each dialect with respect to PC1 and PC2 for each vowel can be seen in Figure 4.6. PCA results for each vowel are discussed in turn.

CHOICE

The first component (PC1) for CHOICE comprises primarily of F2 onset (0.52) and offset (0.44), and Vector Angle (0.50), reflecting variation in the forward-back dimension of the vowel nucleus and glide, as well as the direction of the glide. Such

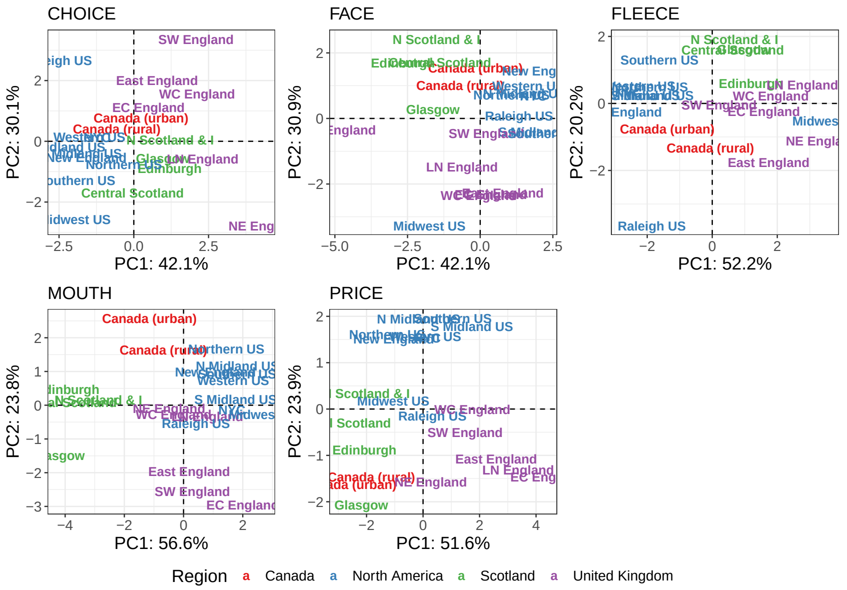

Figure 4.6: PCA biplots showing value of each dialect for the first two principal components for each vowel. Percentages in axes correspond to the proportion of the variance captured by the principal component (Table 4.4, row 10).

variation is observable in Figure 4.2 (outer left), where vowels take on a range of positions in the formant onset. Broadly speaking, variation in this dimension distinguishes North American with generally more backed realisations of CHOICE (both in onset and offset), and British English dialects with fronter CHOICE realisations and more upwards-moving trajectories (Fig. 4.6, outer left). For example, NE England has a very fronted starting point, as well as a distinct upwards-moving inglide (−54.8°, Fig. 4.3); Raleigh, in contrast, has a substantially more backed onset for CHOICE, with a downwards-moving inglide (−96.8°). The second component (PC2) mainly involves F1 onset (−0.46), Vector Length (−0.43), and duration (−0.52). North American and Scottish dialects appear to realise their CHOICE nucleus higher than dialects from England (Fig. 4.2), which may reflect dialect-specific differences in realising CHOICE in a /oɪ/∼/ɔɪ/ spectrum (Wells, 1982). Differences in Vector Length capture some additional resolution regarding dialectal differences in the overall degree of spectral change in the trajectory, mainly reflecting how many dialects broadly converge on a similar glide position in spite of variation in the nucleus. Duration also plays a role in distinguishing dialects in this dimension, where dialects such as Midwest US, Central Scotland, and New England have substantially shorter CHOICE realisations than other dialects.

FACE

The first component (PC1) for FACE comprises primarily of F2 formant positions (onset = -0.51, offset = -0.41) and Vector Offset (-0.45). As can be observed in Figure 4.2, vowels vary substantially in the front-back position of the vowel trajectory. For example, Midwest has a much more fronted FACE vowel, in contrast to New England's more backed realisation. NE England is also exhibits more non-linearity than that other dialects in the realisation of FACE, reflected in a higher Vector Offset value (0.29, Fig. 4.4). The second component (PC2) is comprised predominantly of F1 onset (-0.54), Vector Length (-0.45), and duration (-0.49), which mainly reflects the observation that a range of dialects – Canadian, Scottish, Northern England – have more monophthongal realisations of FACE (Haddican et al., 2013; Wells, 1982). The higher F1 onset position, along with the shorter Vector Length, demonstrates that these dialects realise FACE vowels in a similar position to the endpoint of other dialects' glides. These monophthongal-like FACE vowels are also more likely to be produced with shorter duration than in other dialects (Fig. 4.4).

FLEECE

The first component (PC1) for FLEECE comprises F2 position (onset = 0.44, offset = 0.45) and duration (0.43). Figure 4.2 (centre) shows two distinct clusters of FLEECE realisations in the F2 dimension, where a number of US dialects (e.g., NYC, Southern, South Midland) are realised are more backed than other dialects (Fig. 4.6). The second component (PC2), varying mainly in F1 position (onset = 0.47, offset = 0.67) and Vector Length (-0.45) predominantly captures the distinct realisation for Raleigh (Fig. 4.6), which has lower overall F1 but a greater degree of spectral change, reflected in the higher value for Vector Length.

MOUTH

The first component (PC1) for MOUTH corresponds to variation in F1 offset (0.46), Vector Angle (0.42), and Vector Offset (0.41). As shown in Figure 4.6, Scottish dialects are separated in this dimension, where they exhibit a distinct PRICE-like upgliding trajectory (Fig. 4.2). The second component (PC2), comprising of F2 position (onset = -0.68, offset = -0.41) and Vector Length (0.42), captures the variation in non-Scottish dialects in the overall position and spectral change for the back upgliding diphthong.

PRICE

The first component (PC1) for PRICE comprises mainly of F1 offset (0.47), F2 offset (-0.36), and duration (0.42), which distinguishes between two distinct glide endpoints for PRICE, where British English and Canada dialects have higher and more fronted glide endpoints (Fig. 4.2) and longer durations (Fig. 4.4). The second component (PC2) distinguishes dialects based predominantly on F1 onset (0.46), F2 offset (-0.38), and Vector Angle (0.54): as shown in Figure 4.6, PC2 separates a group of US dialects which are realised with a lower nuclear position, and a steeper glide direction, resulting in a glide that is further back in the F2 dimension: the recognition of Canadian and Scottish dialects with a higher F1 onset may reflect the presence of raising-like patterns in this data (Aitken, 1981; Chambers, 1973).

Summary

The results of the PCA, which enabled inspection of vowel variation across English dialects from all measures together, demonstrate that these vowels differ in structured ways across dialects and across vowels. The dimensions of variation differ for each vowel, but crucially all vowels vary across dialects in terms of both static, dynamic, and duration measures. First, the position of the vowel in formant space is most consistently informative: formant onset or offset appear in the top loadings for all principal components for all five vowels, and F1 or F2 position (onset and offset) account for at least 25% of the variance in all principal components (Table 4.5). Put differently, this means that *no* vowel varies across dialects without *some* degree of variation in overall formant position. Information about the vowel's trajectory shape also accounts for some amount of dialectal variability, where at least one characteristic of trajectory shape appears in the top loadings for one of the first two principal components for each vowel (Table 4.4), and all measures of trajectory shape cumulatively provide at least 25% of the variation for all principal components. Of the trajectory shape measures, Vector Length is most consistently variable across vowels, appearing in the top loadings for all vowels except PRICE. Vector angle is mainly informative for CHOICE and PRICE, and accounts for substantial variation in the principal components for both vowels (25.2% and 29.4% respectively, Table 4.5). Vector offset accounts for the least amount of dialectal variance across vowels, consistent with previous studies examining DCT coefficients (e.g. Williams and Escudero, 2014), and suggests most dialectal variation in trajectory shape concerns linear spectral change. Duration also varies substantially across dialects, playing a role in all vowels except MOUTH.

4.5 Discussion

The role of time-dependent information in the realisation of vowels has been well understood since the earliest acoustic-phonetic analyses of vowels in production (e.g. Joos, 1948; Peterson and Barney, 1952; House, 1961; Gay, 1968). With respect to understanding how time-dependent properties, such as the duration of the vowel and how the vowel changes in its spectral properties over its time-course, differ across vowels, a range of approaches have been applied, such as reporting formant values from multiple points in the vowel (Hillenbrand et al., 1995; Thomas, 2001; Clopper et al., 2005), categorising the presence and direction of a glide (Gay, 1968, 1970; Labov et al., 2006), and formally modelling properties of the formant trajectory itself (e.g. Zahorian and Jagharghi, 1993; Watson and Harrington, 1999; Fox and Jacewicz, 2009; Renwick and Stanley, 2020). Less work has focused on how these representations of dynamic information capture variation across dialects: dialectal differences in vowel duration have predominantly focused on Southern US varieties (Wetzell, 2000; Jacewicz et al., 2007; Tauberer and Evanini, 2009; Fridland et al., 2014), and most studies explicitly modelling dynamic differences across dialects provide comparisons of a small number of closely-related varieties (e.g. Jacewicz et al., 2009; Williams and Escudero, 2014; Swan, 2016; Williams et al., 2019), leaving unaddressed how these time-dependent measures of vowels capture variation across a large number of dialects.

This study provides an exploratory analysis of this question, by considering how measures of formant position, trajectory shape, and duration characterise the dimensions of vowel duration in five vowels across 21 English dialects. First the relative informativity of these measures, alone and together, was formally tested using a dialect classifciation experiment (Section 4.4.1). Then, the variation for these five vowels, within and across the dialects, was examined using dimensionality reduction (Section 4.4.2). It was found that formant position, trajectory shape, and duration together are required to capture dialectal variation in English vowels. Formant position provides the greatest amount of information, reflected in its relatively high performance in dialect classification and its contribution to defining the principal components of all five vowels. Trajectory shape and duration also capture dialectal information in both tasks, but the best performance in dialect classification involved the combination of all measurements. Crucially, the relative informativity of each measure was also shown to differ by vowel class: variation in FLEECE, for example, predominantly involves dialectal differences in the F1 and F2 dimensions, with a substantially smaller role for direct representations of trajectory shape; MOUTH, in contrast, varies in properties of trajectory shape more than in formant position. Together, these results suggest several ways that different properties of vowels can be informative in capturing how dialects systematically differ.

A number of phonetic and sociolinguistic studies have referred to the 'additional resolution' provided by measures that directly represent the shape of the formant trajectory in distinguishing between vowels within the same dialect (e.g. Jacewicz et al., 2009; Williams and Escudero, 2014; Farrington et al., 2018). The results in this study demonstrate that information concerning trajectory shape in fact plays a *primary* role in distinguishing dialects within a given vowel. As illustrated with the dimensionality reduction experiment (Section 4.4.2), at least one measure of trajectory shape provides substantial contribution to the first component for all vowels except PRICE. Results from the dialect classification experiment (Section 4.4.1) show that whilst both formant position and trajectory shape can separately inform the prediction of a given dialect, accuracy is improved with both types of measures are used together. While previous work has shown that trajectory information is informative *within* a given dialect, these results demonstrate that characterisations of the formant trajectory also provide additional resolution as to the ways vowels can systematically differ across individual dialects. This study utilised one particular set of characterisations of trajectory shape – Vector Length, Vector Offset, Vector Angle – and so understanding how other representations of trajectory shape, such as DCTs (Watson and Harrington, 1999; Williams and Escudero, 2014; Williams et al., 2019), would differ in their relative role in distinguishing dialects would be a useful avenue for future research.

Our understanding about cross-dialectal variation in vowel duration has been largely limited to studies of Southern US speech (Jacewicz et al., 2007; Tauberer and Evanini, 2009; Fridland et al., 2014), leaving open the question of how duration varies across English dialects more generally. In this study, a general cross-English effect of duration is observed, though the role of duration appears to be limited to supplementing information provided by formant position and trajectory shape. Within the dialect classification task (Section 4.4.1), predicting the dialect label through the exclusive use of duration as a feature returned the lowest performance of all tests, whilst including duration alongside measures of formant position and trajectory shape resulted in increased performance over classification tests without duration. Duration also played a smaller role in defining the dimensions of variation across dialects, where duration only provided a substantial contribution to the first principal component for FLEECE and PRICE (Section 4.4.2). Given that this study focused on vowels that are typically 'tense' vowels in most English dialects, an expanded analysis including both peripheral and non-peripheral vowel classes (Labov, 1991) would provide additional information about about dialectal differences in duration across English vowels in general.

Analysing data at the scale reported in this study was made possible due to access to a large number of corpora and tools for automated acoustic measurement. Previous large cross-dialectal analyses (e.g. Wells, 1982; Thomas, 2001; Labov et al., 2006) were multi-year enterprises requiring substantial time and labour-intensive manual annotation. Access to force-aligned speech corpora and the automatic measurement of formants allows the analysis to be 'scaled-up' easily relative to many other dialectal studies of vowel quality, but also requires recognition of a number of limitations for studies of this kind. Whilst this method has been shown to generate accurate formant values and procedures are taken to avoid tracking 'false formants' (Mielke et al., 2019), it is simply not possible with data at this scale to be manually validated. Similarly forced alignment have a minimum time duration (often 30ms) and a minimum time resolution (often 10ms), particularly for vowels which may have undergone substantial reduction. We attempted to account for this by applying lower and upper-limits for vowel durations to be included in the study; it remains possible that biases or inaccuracies in vowel duration exist within the dataset.

4.6 Conclusion

Whilst the importance of dynamic time-dependent information has been long-known to play a role in the realisation of vowels, it is unclear how the position of a vowel in formant space, duration, and formant trajectory shape together define how vowels can vary across a large number of dialects. By performing a classification experiment and dimensionality reduction on five vowels in 21 dialects of British and North American English, it was found that all measures of vowels were informative in defining the dimensions of dialectal variation. Formant position provided the greatest degree of informativity, whilst trajectory shape and duration accounted for additional variance. The relative role of each measure was also found to vary on a by-vowel basis, demonstrating that a single set of measures cannot straightforwardly capture the most prominant variation for each vowel.

4.7 Appendices

4.7.1 Parameters for SVM classification experiment

Table 4.6: Optimal values for soft-margin C and kernel γ parameters used in the dialect classification experiment.

	CHOICE		**FACE**		**FLEECE**		**MOUTH**		**PRICE**	
Measures	C	γ	C	γ	C	γ	C	γ	C	γ
Formants (F1, F2 onset + mid + offset)	1000	0.01	1000	0.001	1^{10}	1^{-7}	1	1	1^{6}	0.001
Formants + duration	1^{10}	1^{-7}	1^{5}	0.001	1^{5}	1^{-4}	100	0.01	10	0.1
Duration	1^{10}	1^{-4}	100	1	1^{10}	1^{-6}	1^{7}	1^{-5}	1^{6}	1^{-6}
Trajectory (Vector Length, offset, angle)	10000	0.01	100	0.01	1^{9}	0.001	1^{7}	0.001	1	1
Trajectory + duration	1^{5}	1^{-4}	1000	0.10	10	0.01	1^{5}	0.001	1^{5}	0.001
Formants + Trajectory	1	0.1	1^{7}	1^{-6}	1^{7}	1^{-7}	10000	0.001	1^{6}	1^{-4}
Formants + Trajectory + duration	1^{7}	1^{-6}	1^{6}	1^{-4}	1^{5}	1^{-4}	1^{7}	1^{-4}	1^{5}	0.001

Chapter 5

Conclusion

The goal of this thesis has been to address how phonetic variability is structured across dialects and speakers of two languages – English and Japanese – in three large-scale phonetic studies, characterised by the use of large speech corpora, automatic acoustic measurements, and a quantitative approach to modelling speech variability. The three studies reported in this thesis have explored the realisation of phonological contrasts that exhibit systematic variability at the levels of dialects and individual speakers, where much of this variability is constrained in its scope and context. Section 5.1 summarises the key motivations and results of each study reported in this thesis, Section 5.2 presents a general discussion concerning how these results relate to previous findings and future directions for research on structured variation, and Section 5.3 concludes the thesis.

5.1 Summary

The first study, presented in Chapter 2, examined how speakers varied across two cues to the word-initial stop voicing contrast – Voice Onset Time (VOT) and closure voicing – in a corpus of spontaneous Japanese speech. Previous studies of speaker variation in stop contrasts has focused on Germanic languages (like English and German) in tightly-controlled laboratory speech contexts (e.g. Chodroff and Wilson, 2017, 2018; Hullebus et al., 2018), leaving unclear how speaker variation in stop contrast production may differ in spontaneous speech and in a language with a different phonetic implementation of stop voicing (Nasukawa, 2005). Modelling speaker variation both *within* and *across* both acoustic cues to the stop contrast, it was found that speakers were highly correlated in the use of each cue in isolation to mark the stop voicing contrast: for example, speakers vary in their overall use of VOT, but relative difference in the use of VOT for voiced and voiceless stops is highly constrained. These relationships were much weaker *across* acoustic cues, suggesting that speakers were able to independently vary in their relative use of VOT and closure voicing for marking the voicing contrast. These findings indicate the presence of structured speaker variability in the realisation of the Japanese stop voicing contrast, but also that this variability is itself constrained: compared with robust cross-cue contrasts observed in English (Chodroff and Wilson, 2018; Sonderegger et al., 2020a), the findings in this study suggest that structured variability may also be constrained by the language-specific implementation of linguistic contrasts.

The second study, presented in Chapter 3, examined the degree of variability in the English pre-consonantal voicing effect – the vowel duration difference preceding voiced and voiceless consonants – across dialects and individual speakers. Beyond studies describing a large and perceptually-salient voicing effect for English relative to other languages (House and Fairbanks, 1953; Denes, 1955; Chen, 1970; Raphael, 1972), little is known about variability of the voicing effect internal to a particular language, including how it is modulated by phonetic factors (such as speech rate and word frequency) in spontaneous speech and the extent of its variability across dialects and speakers. Modelling the size of the voicing effect across 30 English dialects, it was found that the overall ('English-wide') voicing effect was substantially smaller in spontaneous speech, as compared with previous reports of laboratory speech, and that the size of the effect was reduced in contexts where the overall vowel duration was shorter (cf. Klatt, 1973). The voicing effect was also shown to be highly variable across dialects of English, ranging from near-null sizes for Scottish varieties to African American varieties exhibiting the largest sizes. Speakers varied less than dialects, suggesting that individual speakers likely deviate little from their dialect-specific baseline.

The third study, presented in Chapter 4, investigated how multiple time-dependent acoustic cues to vowels – formant position, trajectory shape, duration – define the ways in which vowels vary across a large number of English dialects. Whilst time-dependent variation in vowels has been long-recognised and studied within individual dialects (e.g. Peterson and Barney, 1952; Hillenbrand et al., 1995; Watson and Harrington, 1999), much cross-dialectal research on vowels has focused on static properties (Labov et al., 1972; Thomas, 2001; Labov et al., 2006) or on a small number of closely-related dialects (e.g. Jacewicz et al., 2009; Williams and Escudero, 2014; Farrington et al., 2018). In this study, data from five vowels across 21 English dialects were examined. It was found that information regarding the vowel's onset and offset in formant space was highly informative as a measure of dialectal variation, followed by measures directly corresponding to the shape of the formant trajectory. Duration was less informative as a singular cue, but played an important supplementary role to measures of formant position and shape. Dialectal variability was also shown to be highly structured across these measures, where the majority of dialectal variation for any given vowel could be explained in terms of linear combinations of these time-dependent properties. These results demonstrate the importance of considering time-dependent dynamic information in characterising how vowels can vary across dialects of a language, and indicate that dialectal variation in vowels is highly structured.

5.2 General discussion

5.2.1 Structured variability

Understanding the sources and structure of phonetic variation have been central themes in phonetic and sociolinguistic research: in contrast to variability being random and unstructured, phonetic variation exhibits underlying *structure* which is explainable as a function of a number of linguistic, social, and cognitive factors. This thesis has focused on two domains of speech variability: variability across *dialects* and across *speakers.*

Chapters 2 and 3 explored the structure of variability across speakers, and observed that individual speakers exhibit substantial uniformity in the realisation of voicing contrasts. Such constraints on speaker variation are likely useful for speech perception: speech variability is highly multi-dimensional (Liberman et al., 1967), and constraints on the ways in which speakers can differ creates a lower-dimensional space for speaker variation, likely aiding in speaker-independent normalisation in perception (Creel and Bregman, 2011; Trude and Brown-Schmidt, 2012; Chodroff, 2017; Chodroff and Wilson, 2017, 2018; Kleinschmidt, 2018). Similarly, underlying structure in speaker variability is also likely defined with respect to community-level patterns (e.g. Wolfram and Beckett, 2000; Schilling-Estes, 2004; Labov, 2014), where such constraints may also contribute to why some phonetic processes remain relatively stable across time whilst others undergo diachronic change (Weinreich et al., 1968; Ohala, 1989; Baker et al., 2011).

Chapters 3 and 4 provide further evidence regarding the structure of variation across dialects of the same language. Dialects were constrained as to the range of possible voicing effect sizes, whilst the majority of variability in vowel realisation could be expressed in lower-dimensional combinations of acoustic cues. These findings demonstrate that languages can maintain a reasonable degree of internal variation (e.g., across dialects), but there exist limits to dialectal variation, both in terms of *what* can vary and in *how* much a given dialect can deviate from other varieties. These constraints may be functional in nature; as is the case with speaker-level variability, dialects may be limited in their ability to deviate from each other due to some demand of mutual intelligibility, or perhaps due to limits on the phonetic implementation of otherwise similar phonological structures (Keating, 1985). For example, whilst languages may differ in the phonological specification of laryngeal contrasts (Iverson and Salmons, 1995; Beckman et al., 2013; Salmons, 2019), dialects of a particular language may only be limited to determining *how* that particular phonological specification is mapped into acoustic-phonetic implementation.

5.2.2 'Large-scale' methodologies

The conclusions that these studies have been able to draw about dialect and speaker variability were made possible by the ability to collect and process speech data at a relatively large scale, relative to traditional studies within phonetic and sociolinguistic research. A handful of projects have systematically collected and analysed phonetic data from a large number of dialects (Thomas, 2001; Labov et al., 2006), but the methodological challenges of such projects – sourcing and recording speakers, transcribing and aligning speech data, performing acoustic analyses – constrain the research to be performed at the level of single or a small number of closely-related speech communities. This was made possible in this thesis through the use of already-collected speech corpora: for Chapter 2, this refers to a publically-available Japanese speech corpus (Maekawa et al., 2000). For Chapters 3 and 4, access and processing of speech corpora was made possible via data-sharing in the SPADE project (Sonderegger et al., 2020b).

Whilst the collection of large amounts of speech data was comparatively straightforward within this thesis due to its prior availability, the size of the resulting data made it so that manual acoustic measurement would have been too time and labour-intensive to be feasible. Instead, processing this data was made possible through the use of automatic tools: forced alignment (e.g. Rosenfelder et al., 2014; Gorman et al., 2011; McAuliffe et al., 2017a), the subsequent importation and processing heterogeneous corpus formats with the ISCAN tool (McAuliffe et al., 2019) and acoustic measurements of voicing (Boersma and Weenink, 2017) and vowel duration and formants (Mielke et al., 2019). Whilst these tools vastly increase the number and ease by which acoustic measurements can be made, they also increase the number of measurement errors that may be introduced into datasets. As the number of observations in these datasets is too large to manually correct, this thesis applied a conservative approach in deciding the quality of measurements. For example, vowels with durations either at or close to the minimum durations from forced alignment usually indicate a case of poor alignment, and such tokens are discarded from analyses. The approach to statistical modelling taken in Chapters 2 and 3 explicitly estimate both the size of a particular effect of interest (e.g., a vowel's duration in a given context), but also the range of likely values given the data and the model: this makes it possible to quantify the uncertainty associated with a particular effect in a particular dataset, which provides some protection from over-interpreting the results within a given study (Vasishth et al., 2018a).

5.2.3 Future directions

Although structured variation has been at the centre of phonetic and sociolinguistic research for more than sixty years, there still remain many unknowns about the sources and structure of phonetic variation across dialects and speakers. As evidenced in Chapter 3, even our understanding of a relatively well-studied phonetic variable like the English voicing effect is improved through characterising its variability at multiple levels of linguistic structure. Both this thesis and other recent research on structured variability have focused on how speakers vary in the perception and realisation of phonological contrasts (e.g. Schultz et al., 2012; Schertz et al., 2015; Clayards, 2018b; Chodroff and Wilson, 2017), whilst dialectological and sociolinguistic studies of dialect variation have often explored non-phonemic variation such as coronal stop deletion (e.g. Guy, 1980; Tagliamonte and Temple, 2005; Hazen, 2011). A potentially-promising avenue of research would concern the scope and distribution of speaker variability for non-phonemic contrasts, which may provide information as to how speakers may differ in their specific phonological and phonetic implementation. Understanding whether individual speakers differ in their deletion rate or are differently influenced by modulating factors on deletion could point to instances of dynamic change within speech communities not directly tied to phonemic structure. Another direction could consider extending the analysis of structured variability to cross-linguistic variation, which may inform theories concerning what elements of phonetic structure may be implemented in a language-specific fashion, or how languages with different *phonological* structures share similar *phonetic* constraints (Keating, 1985; Chodroff and Wilson, 2017).

This thesis has also demonstrated the scientific utility of modelling phonetic variability across a large number of dialects and speakers, and it seems reasonable to expect that many future studies within the phonetic and sociolinguistic paradigms will further utilise this ‘large-scale’ corpus approach (Liberman, 2018). Given increased access to speech corpora and tools for processing and measuring speech, analyses using large-scale data provide an opportunity to develop and test linguistic theories in an ‘ecologically-valid’ way – examining speech in a naturalistic context. Finally, an integrated approach to corpus analysis – like that exemplified by SPADE and ISCAN – could be utilised for cross-linguistic research.

5.3 Conclusion

The three studies reported in this thesis have explored the sources and structure of phonetic variability across dialects and speakers. It was found in all three studies that, far from being random, phonetic variation can be characterised with respect to linguistic and social factors. By applying a 'large-scale' approach to the study of structured variability, this thesis has demonstrated the value in examining speech variability across a wide number of dialects and speakers for developing theories of phonetic and linguistic structure.

Bibliography

Abramson, A. S. and Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, 63:75–86.

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

Aitken, A. J. (1981). *The Scottish Vowel Length Rule*. The Middle English Dialect Project, Edinburgh.

Allen, S. J., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 113:544–552.

Anderson, J., Beavan, D., and Kay, C. (2007). The Scottish corpus of texts and speech. In Beal, J. C., Corrigan, K. P., and Moisl, H. L., editors, *Creating and Digitizing Language Corpora*, pages 17–34. Palgrave, New York.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America*, 71:967–989.

Bailey, C.-J. (1968). Segmental length in Southern States English: an instrumental phonetic representation of a standard dialect in South Carolina. In *PEGS Paper No. 20*. Center for Applied Linguistics, Washington DC.

Baker, A., Archangeli, D., and Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change*, 23.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Bang, H.-Y. (2017). *The structure of multiple cues to stop categorization and its implications for sound change*. PhD thesis, McGill University.

Baran, J., Laufer, M., and Daniloff, R. (1977). Phonological contrastivity in conversation: a comparative study of voice onset time. *Journal of Phonetics*, 5:339–350.

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

Barr, D. J., Levy, R., Sheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

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

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

Beddor, P., Coetzee, A., Styler, W., and McGowan, K. (2018). The time course of individuals perception of coarticulatory information is linked to their production: Implications for sound change. *Language*, 94:931–968.

Bekker, I. (2012). The story of South African English: a brief linguistic overview. *International Journal of Language, Translation & Intercultural Communication*, 1:139–150.

Bell, A. (1984). Language style as audience design. *Language in Society*, 12:145–204.

Bird, S. and Liberman, M. (1999). Annotation graphs as a framework for multidimensional linguistic data analysis. In *Towards Standards and Tools for Discourse Tagging*.

Boberg, C. (2008). Regional phonetic variation in Standard Canadian speech. *Journal of English Linguistics*, 36:129–154.

Boberg, C. (2010). *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge University Press, Cambridge.

Boberg, C. (2018). Dialects of North American English. In Boberg, C., Nerbonne, J., and Watt, D., editors, *Handbook of Dialectology*, pages 450–461. John Wiley and Sons.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer (version 6.0.36).

Bois, J. W. D., Chafe, W. L., Meyer, S. A., Thompson, S. A., and Martey, N. (2000). Santa Barbara corpus of Spoken American English. Technical report, Linguistic Data Consortium, Philadelphia.

Bradley, D. (2004). Regional characteristics of Australian English: phonology. In Schneider, E., Burridge, K., Kortmann, B., Mesthrie, R., and Upton, C., editors, *A Handbook of Varieties of English: Volume 1 Phonology*. Mouton de Gruyter, New York.

Browman, C. P. and Goldstein, L. (1991). Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech.*, pages 341–376. Cambridge University Press.

Brysbaert, M. and 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. *Behavioral Research Methods*, 41:977–990.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Bybee, J. B. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge.

Byrd, D. (1993). 54,000 American stops. *UCLA Working Papers in Linguistics*, 83:97–116.

Cardoso, A. B. (2015). *Dialectology, phonology, diachrony: Liverpool English realisations of price and mouth*. PhD thesis, University of Edinburgh.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Chambers, J. K. (1973). Canadian Raising. *Canadian Journal of Linguistics*, 18:113–135.

Chambers, J. K. and Trudgill, P. (1980). *Dialectology*. Cambridge University Press, Cambridge.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22:129–159.

Chiba, T. and Kajiyama, M. (1941). *The Vowel: Its Nature and Structure*. Kaiseikan, Tokyo.

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

Cho, T. and McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, 33:121–157.

Chodroff, E. (2017). *Structured Variation in Obstruent Production and Perception*. PhD thesis, Johns Hopkins University.

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

Chodroff, E. and Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, 4.

Clarke, S., Elms, F., and Youssef, A. (1995). The third dialect of English: some Canadian evidence. *Language Variation and Change*, 7:209–228.

Clayards, M. (2018a). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *Journal of the Acoustical Society of America*, 144:EL172–177.

Clayards, M. (2018b). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, 75:1–23.

Clopper, C. G. (2009). Computational methods for normalizing acoustic vowel data for talker difference.

Clopper, C. G., Pisoni, D. B., and de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America*, 118:1661–1676.

Clopper, C. G. and Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in american english. *Journal of Phonetics*, 39(2):237–245.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ.

Cohen Priva, U. and Gleason, E. (2018). The role of fast speech in sound change. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 1512–1517, Austin, TX. Cognitive Science Society.

Cole, A. and Strycharczuk, P. (2019). The PRICE-MOUTH crossover in the "Cockney diaspora". In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). Audio BNC: the audio edition of the Spoken British National Corpus. Technical report, Oxford. http://www.phon.ox.ac.uk/AudioBNC.

Coleman, J., Renwick, M. E. L., and Temple, R. A. M. (2016). Probabilistic underspecification in nasal place assimilation. *Phonology*, 33(3):425458.

Coretta, S. (2019). An exploratory study of voicing-related differences in vowel duration as compensatory temporal adjustment in Italian and Polish. *Glossa: A Journal of General Linguistics*, 4:1–25.

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

Crystal, T. H. and House, A. S. (1982). Segmental durations in connected speech signals: preliminary results. *Journal of the Acoustical Society of America*, 72:705–716.

Crystal, T. H. and House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, 88:101–112.

Cuartero, N. (2002). *Voicing assimilation in Catalan and English*. PhD thesis, Universitat Autònoma de Barcelona.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, 54:35–60.

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

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 27:761–764.

Deterding, D. (2003). An instrumental study of the monophtong vowels of Singapore English. *English World-Wide*, 24:1–16.

DiCanio, C., Nam, H., Amith, J. A., Garcia, R. C., and Whalen, D. H. (2015). Vowel variablility in elicited versus spontaneous speech: evidence from Mixtec. *Journal of Phonetics*, 48:45–59.

Docherty, G. (1992). *The timing of voicing in British English obstruents*. Foris, Berlin & New York.

Docherty, G., Gonzalez, S., and Mitchell, N. (2015). Static vs dynamic perspectives on the realization of vowel nucleii in West Australian English. In *Proceedings of the 18th International Congress of Phonetic Sciences*.

Dodsworth, R. (2013). Retreat from the Southern Vowel Shift in Raleigh, NC: social factors. *University of Pennsylvania Working Papers in Linguistics*, 19:31–40.

Dodsworth, R. and Kohn, M. (2012). Urban rejection of the vernacular: The SVS undone. *Language Variation and Change*, 24:221–245.

Eager, C. (2015). Automated voicing analysis in Praat: statistically equivalent to manual segmentation. In *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Eckert, P. (2012). Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41:87–100.

Eckert, P. (2019). The limits of meaning. *Language*, 95:87–100.

Ernestus, M., Hanique, I., and Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *Journal of Phonetics*, 38:60–75.

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

Evans, M. (1935). Southern 'long i'. *American Speech*, 10:188–190.

Fabricius, A. H. (2000). *T-glottalling between stigma and prestige: a sociolinguistic study of Modern RP*. PhD thesis, Copenhagen Business School, Copenhagen, Denmark.

Fant, G. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. In Halle, M., Lunt, H. G., and Schooneveld, C. V., editors, *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday*. Mouton, The Hague.

Farrington, C. (2018). Incomplete neutralization in African American English: the case of final consonant devoicing. *Language Variation and Change*, 30:361–383.

Farrington, C., Kendall, T., and Fridland, V. (2018). Vowel dynamics in the southern vowel shift. *American Speech*, 93:186–222.

Fitt, S. (2001). *Unisyn Lexicon*. Centre for Speech Technology Research, Edinburgh.

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

Foulkes, P. and Docherty, G. (1999). *Urban Voices: Accent studies in the British Isles*. Edward Arnold.

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

Foulkes, P., Docherty, G., and Watt, D. (2001). The emergence of structured variation. *University of Pennsylvania Working Papers in Linguistics*, 7:67–84.

Foulkes, P., Docherty, G., and Watt, D. (2005). Phonological variation in child-directed speech. *Language*, 81:177–206.

Foulkes, P., Scobbie, J., and Watt, D. (2010). Sociophonetics. In Hardcastle, W., Laver, J., and Gibbon, F., editors, *The Handbook of Phonetic Sciences*, pages 703–754. Blackwell, Oxford.

Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in English. *Language and Speech*, 26:21–60.

Fox, R. A. and Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English. *Journal of the Acoustical Society of America*, 126:2603–2618.

Fridland, V. (2000). The Southern shift in Memphis. *Language Variation and Change*, 11:267–285.

Fridland, V., Kendall, T., and Farrington, C. (2014). Durational and spectral differences in American English vowels: dialect variation within and across groups. *Journal of the Acoustical Society of America*, 136:341–349.

Fromkin, V. A. (1977). Some questions regarding universal phonetics and phonetic representations. In Juilland, A., editor, *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday*, pages 365–380. Anma Libri, Saratoga.

Fromont, R. and Hay, J. (2012). LaBB-CAT: an annotation store. In *Australasian Language Technology Workshop 2012*, volume 113, pages 113–117.

Fruehwald, J. (2013). *The Phonological Influence on Phonetic Change*. PhD thesis, University of Pennsylvania.

Fruehwald, J. (2016a). The early influence of phonology on a phonetic change. *Language*, 92:376–410.

Fruehwald, J. (2016b). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, 22:41–49.

Fujimoto, M., Kikuchi, H., and Maekawa, K. (2006). Corpus of Spontaneous Japanese documentation: phone information. Technical Report 6, National Institute for Japanese Language and Linguistics, Tokyo.

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

Gao, J. and Arai, T. (2019). Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics*, 77:1–33.

Gao, J., Yun, J., and Arai, T. (2019). VOT and F0 coarticulation in Japanese: production-biased or misparsing? In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). Darpa timit acoustic phonetic continuous speech corpus cdrom.

Gay, T. (1968). Effects of speaking rate on dipthong formant movements. *Journal of the Acoustical Society of America*, 44:1570–1573.

Gay, T. (1970). A perceptual study of American English diphthongs. *Language & Speech*, 13:65–88.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge.

Gendrot, C. and Adda-Decker, M. (2005). Impact on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Interspeech 2005*, pages 2453–2456.

Geng, C., Turk, A., Scobbie, J. M., Macmartin, C., Hoole, P., Richmond, K., Wrench, A., Pouplier, M., Bard, E. G., Campbell, Z., Dickie, C., Dubourg, E., Hardcastle, W., Kainada, E., King, S., Lickley, R., Nakai, S., Renals, S., White, K., and Wiegand, R. (2013). Recording speech articulation in dialogue: Evaluating a synchronized double electromagnetic articulography setup. *Journal of Phonetics*, 41:421–431.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, pages 517–520.

Gordon, E., Maclagan, M., and Hay, J. (2007). The ONZE corpus. In Beal, J., Corrigan, K., and Moisl, H., editors, *Creating and Digitizing Language Corpora*. Palgrave-Macmillan, New York.

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

Grabe, E. (2004). Intonational variation in English. In Gilles, P. and Peters, J., editors, *Regional Variation in Intonation*, pages 9–31. Niemeyer, Tubingen.

Greenbaum, S. and Nelson, G. (1996). The International Corpus of English (ICE project). *World Englishes*, 15:3–15.

Gubian, M., Torreira, F., and Boves, L. (2015). Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49:16–40.

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

Haddican, B., Foulkes, P., Hughes, V., and Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change*, 25:371–403.

Harris, M. and Umeda, N. (1974). Effect of speaking mode on temporal factors in speech: vowel duration. *The Journal of the Acoustical Society of America*, 56(3):1016–1018.

Hauser, I. (2019). *Effects of Phonological Contrast on Within-Category Phonetic Variation*. PhD thesis, University of Massachusetts Amherst.

Hay, J. and Drager, K. (2007). Sociophonetics. *Annual Review of Anthropology*, 36:89–103.

Hazen, K. (2011). Flying high above the social radar: Coronal stop deletion in modern Appalachia. *Language Variation and Change*, 23:105–137.

Heffner, R.-M. S. (1937). Notes on the lengths of vowels. *American Speech*, 12:128–134.

Hewlett, N., Matthews, B., and Scobbie, J. M. (1999). Vowel duration in Scottish English speaking children. In *Proceedings of 14th The International Congress of Phonetic Sciences*, San Francisco.

Hibbitt, G. W. (1948). *Diphthongs in American speech : a study of the duration of diphthongs in the contextual speech of two hundred and ten male undergraduates*. PhD thesis, Columbia University.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of English vowels. *Journal of the Acoustical Society of America*, 97:3099–3111.

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (2001). Effect of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109:748–763.

Holmes-Elliott, S. (2015). *London calling: assessing the spread of metropolitan features in the southeast*. PhD thesis, University of Glasgow.

Holt, Y. F. and Ellis, C. (2018). African American women's speech: vowel inherent spectral change. *Acoustic Science & Technology*, 39:160–162.

Holt, Y. F., Jacewicz, E., and Fox, R. A. (2016). Temporal variation in African American English: the distinctive use of vowel duration. *Journal of Phonetics & Audiology*, 2.

House, A. S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, 33:1174–1178.

House, A. S. and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 25:105–113.

Hullebus, M. A., Tobin, S. J., and Gafos, A. I. (2018). Speaker-specific structure in German voiceless stop voice onset times. In *Proceedings of Interspeech 2018*, pages 1403–1407, Hyderabad.

Hunnicutt, L. and Morris, P. A. (2016). Prevoicing and aspiration in Southern American English. In *Proceedings of the 39th Annual Penn Linguistics Conference*, volume 22, Philadelphia. University of Pennsylvania.

Ito, J. and Mester, A. R. (1995). Japanese phonology. In Goldsmith, J. A., editor, *The Handbook of Phonological Theory*, pages 817–838. Blackwell.

Iverson, G. and Salmons, J. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, 12:369–396.

Jacewicz, E. and Fox, R. A. (2013). Cross-dialectal differences in dynamic formant patterns in American English vowels. In Morrison, G. S. and Assmann, P. F., editors, *Vowel Inherent Spectral Change*, pages 177–198. Springer, Berlin.

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

Jacewicz, E., Fox, R. A., and Salmons, J. (2007). Vowel duration in three American English dialects. *American Speech*, 82:367–385.

Johnson, K. (2004). Massive reduction in conversational American English. In Yoneyama, K. and Maekawa, K., editors, *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, pages 29–54. The National International Institute for Japanese Language, Tokyo.

Johnson, K., Ladefoged, P., and Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America*, 94:701–714.

Johnstone, B. (1996). *The Linguistic Individual: self-expression in language and linguistics*. Oxford University Press, Oxford.

Johnstone, B., Bashin, N., and Wittkofski, D. (2002). 'Dahntahn' Pittsburgh: monophthongal /aw/ and representations of localness in Southwestern Pennsylvania. *American Speech*, 77:148–166.

Johnstone, B., Baumgardt, D., Eberhardt, M., and Kiesling, S. (2015). *Pittsburgh Speech and Pittsburghese*. Walter de Gruyter, Berlin.

Jones, D. (1909). *Intonation Curves: A Collection of Phonetic Texts, in Which Intonation Is Marked Throughout by Means of Curved Lines on a Musical Stave*. Teubner, Berline.

Jones, D. (1948). *An Outline of English phonetics*. E. P. Dutton & Company, New York.

Joos, M. (1948). Acoustic phonetics. *Language*, 24:1–136.

Kawahara, S. (2015). Geminate devoicing in Japanese loanwords: Theoretical and experimental investigations. *Language and Linguistics Compass*, 9:181–195.

Kay, M. (2019). *tidybayes: Tidy Data and Geoms for Bayesian Models*. R package version 1.0.4.

Keating, P. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, 60:189–218.

Keating, P. (2006). Phonetic encoding of prosodic structure. In Harrington, J. and Tabain, M., editors, *Speech Production: Models, phonetic processes, and techniques*, pages 197–186. Psychology Press, New York.

Keating, P., Byrd, D., Flemming, E., and Todaka, Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication*, 14:131–142.

Keating, P. A. (1985). Universal phonetics and the organization of grammars. In Fromkin, V. A., editor, *Phonetic Linguistics: essays in honor of Peter Ladefoged*, pages 115–132. Academic Press, New York.

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Fundamental of Machine Learning for Predictive Data Analytics*. MIT Press, Cambridge MA.

Kendall, T. (2007). The sociolinguistic archive and analysis project: Empowering the sociolinguistic archive. *Pennsylvania Working Papers in Linguistics*, 13:15–26.

Kendall, T. (2013). *Speech rate, Pause, and Socioliguistic Variation*. Palgrave MacMillan, London.

Kendall, T. and Farrington, C. (2018). The Corpus of Regional African American Language. Version 2018.10.06.

Kenyon, J. S. (1940). *American Pronunciation*. George Wahr, Ann Arbor.

Kerswill, P., Torgersen, E. N., and Fox, S. (2008). Reversing 'drift': Innovation and diffusion in the London diphthong system. *Language Variation and Change*, 20:451–491.

Kikuchi, H. and Maekawa, K. (2003). Performance of segmental and prosodic labeling of spontaneous speech. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo. Tokyo Institute of Technology.

Kim, D. and Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition & Neuroscience*, 34:769–786.

Kim, S., Kim, J., and Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, 71:65–80.

Kirkham, S., Nance, C., Littlewood, B., Lightfoot, K., and Groake, E. (2019). Dialect variation in formant dynamics: the acoustics of lateral and vowel sequences in Manchester and Liverpool English. *Journal of the Acoustical Society of America*, 145:784–794.

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

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, 54:1102–1104.

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.

Kleber, F. (2018). VOT or quantity: What matters more for the voicing contrast in German regional varieties? results from apparent-time analyses. *Journal of Phonetics*, 71:468–486.

Kleinschmidt, D. F. (2018). Structure in talker variability: How much is there and how much can it help? *Language, Cognition, and Neuroscience*, 34:1–26.

Knowles, G. (1973). *Scouse: the urban dialect of Liverpool*. PhD thesis, University of Leeds.

Koenig, W., Dunn, H. K., and Lacy, L. Y. (1946). The sound spectrograph. *Journal of the Acoustical Society of America*, 18:19–49.

Kong, E. J. and Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, 59:40–57.

Kong, E. J., Yoneyama, K., and Beckman, M. E. (2014). Effects of a sound change in progress on gender-marking cues in Japanese. In *Proceedings of LabPhon 14*, Tokyo. National Institute for Japanese Language and Linguistics.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York.

Kuhn, M. and Vaughan, D. (2020). *yardstick: Tidy Characterizations of Model Performance*. R package version 0.0.6.

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

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington DC.

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

Labov, W. (1991). Three dialects of English. In Eckert, P., editor, *New ways of analyzing variation in English*, pages 1–45. Academic, New York.

Labov, W. (1994). *Principles of Language Change, Vol 1: Internal Factors*. Blackwell, Malden, MA.

Labov, W. (2001). *Principles of Linguistic Change, Vol 2: Social Factors*. Blackwell, Malden, MA.

Labov, W. (2011). *Principles of Language Change, Vol 3: Cognitive and Cultural Factors*. Blackwell, Malden, MA.

Labov, W. (2014). The sociophonetic orientation of the language learner. In Celeta, C. and Calamai, S., editors, *Advances in Sociophonetics*, chapter 1, pages 17–30. John Benjamins, Amsterdam.

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Mouton de Gruyter, Berlin.

Labov, W., Cohen, P., Robins, C., and Lewis, J. (1968). A study of the non-standard english of Negro and Puerto Rican speakers in New York City. Technical Report 1 & 2, Linguistics Laboratory, University of Pennsylvania.

Labov, W. and Rosenfelder, I. (2011a). New tools and methods for very large scale measurements of very large corpora. In *New Tools and Methods for Very-Large-Scale Phonetics Research Workshop*.

Labov, W. and Rosenfelder, I. (2011b). *The Philadelphia Neighborhood Corpus*. University of Pennsylvania Linguistics Laboratory, Philadelphia.

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language*, 89:30–65.

Labov, W., Yaeger, M., and Steiner, R. (1972). *A quantitative study of sound change in progress: Volume 1. Report on National Science Foundation Contract NSF-GS-3287*. University of Philadelphia, Pennsylvania.

Ladefoged, P. and Maddieson, I. (1993). *The Sounds of the World's Languages*. Wiley, Oxford.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Lehiste, I. (1970). Temporal organization of higher-level linguistic units. *Journal of the Acoustical Society of America*, 48:111.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100:1989–2001.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1:153–167.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–369.

Liberman, M. (2018). Corpus phonetics. *Annual Review of Linguistics*, 5:91–107.

Liberman, M. A., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74:431–461.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, volume 4 of *NATO ASI Series*, pages 403–439. Kluwer Academic Publishers.

Lisker, L. (1957). Linguistic segments, acoustic segments and synthetic speech. *Language*, 33:370–374.

Lisker, L. (1985). The pursuit of invariance in speech signals. *Journal of the Acoustical Society of America*, 77:1199–1202.

Lisker, L. (1986). Voicing in English: a catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech*, 29:3–11.

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

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

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, 49:606–608.

Luce, P. A. and Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, 78:1949–1957.

Macaulay, R. K. S. (1977). *Language, social class and education*. University of Edinburgh Press, Edinburgh.

Mack, M. (1982). Voicing-dependent vowel duration in English and French: monolingual and bilingual production. *Journal of the Acoustical Society of America*, 71:173–178.

Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: An extended J_ToBI for spontaneous speech. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 1545–1548, Denver.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference of Language Resources and Evaluation (LREC)*, volume 2, pages 946–952.

Magnuson, J. S. and Nusbaum, H. C. (2007). Acoustic difference, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology*, 33:391–409.

Mayr, R. and Davies, H. (2011). A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh. *Journal of the International Phonetic Association*, 41:1–25.

McAuliffe, M., Coles, A., Goodale, M., Mihuc, S., Wagner, M., Stuart-Smith, J., and Sonderegger, M. (2019). ISCAN: A system for integrated phonetic analyses across speech corpora. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne.

McAuliffe, M., Scolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017a). Montreal forced aligner [computer program]. https://montrealcorpustools.github.io/Montreal-Forced-Aligner/.

McAuliffe, M., Stengel-Eskin, E., Socolof, M., and Sonderegger, M. (2017b). Polyglot and Speech Corpus Tools: a system for representing, integrating, and querying speech corpora. In *Proceedings of Interspeech 2017*.

Mendoza-Denton, N. (2010). Individuals and communities. In Wodak, R., Johnstone, B., and Kerswill, P., editors, *The SAGE Handbook of Sociolinguistics*, pages 181–191. SAGE, London.

Mester, A. and Ito, J. (1989). Feature predictability and underspecification: palatal prosody in Japanese mimetics. *Language*, 65:258–293.

Meunier, C. and Espresser, R. (2011). Vowel reduction in casual French: the role of lexical factors. *Journal of Phonetics*, 39:271–278.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-3.

Mielke, J., Baker, A., and Archangeli, D. (2016). Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/. *Language*, 92:101–140.

Mielke, J., Thomas, E. R., Fruehwald, J., McAuliffe, M., Sonderegger, M., Stuart-Smith, J., and Dodsworth, R. (2019). Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several English speech corpora. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Morrison, G. S. (2013). Theories of vowel inherent spectral change. In Morrison, G. S. and Assman, P. F., editors, *Vowel Inherent Spectral Change*, pages 49–86. Springer, Berlin.

Nasukawa, K. (2005). The representation of laryngeal-source contrasts in Japanese. In van de Weijer, J., Nanjo, K., and Nishihara, T. ., editors, *Voicing in Japanese*, pages 71–87. De Gruyter Mouton.

Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*. PhD thesis, Indiana University Linguistics Club.

Nearey, T. M. and Assmann, P. F. (1986). Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, 80:1297–1308.

Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – part II. *Language and Linguistics Compass*, 10:591–613.

Ohala, J. (1989). Sound change is drawn from a pool of synchronic variation. In Breivik, L. E. and Jahr, E. H., editors, *Language Change: Contributions to the study of its causes*, pages 173–198. Mouton de Gruyter, Berlin.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184.

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

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lentition, and contrast. In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*, pages 137–157. John Benjamins.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). *Buckeye Corpus of Spontaneous Speech*. Ohio State University, Columbus, 2 edition.

Port, R. F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, 32:141–152.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. In Pisoni, D. B. and Remez, R. E., editors, *The Handbook of Speech Perception*, pages 182–206. Blackwell, Oxford.

Rathcke, T. and Stuart-Smith, J. (2016). On the tail of the Scottish Vowel Length Rule in Glasgow. *Language and Speech*, 59:404–430.

Renwick, M. E. L. and Stanley, J. A. (2020). Modeling dynamic trajectories of front vowels in the American South. *Journal of the Acoustical Society of America*, 147:579–595.

Riney, T. J., Takagi, N., Ota, K., and Uchida, Y. (2007). The intermediate degree of VOT in Japanese initial stops. *Journal of Phonetics*, 35:439–443.

Risdal, M. L. and Kohn, M. E. (2014). Ethnolectal and generational differences in vowel trajectories: Evidence from African American English and the Southern vowel system. In *Selected papers from NWAV 42*, pages 139–148, Philadelphia. University of Pennsylvania.

Roettger, T. B., Winter, B., and Baayen, R. H. (2019). Emergent data analysis in phonetic sciences: towards pluralism and reproducibility. *Journal of Phonetics*, 73:1–7.

Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., and Wareham, T. (2006). Introducing Phon: a software solution for the study of phonological acquisition. In *Proceedings of the 30th Annual Boston University Conference on Language Development*, pages 489–500.

Rosen, N. and Skriver, C. (2015). Vowel patterning of Mormons in Southern Alberta, Canada. *Language & Communication*, 42:104–115.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., and Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) program suite v1.2.2 10.5281/zenodo.22281.

Rositzke, H. A. (1939). Vowel-length in General American speech. *American Speech*, 15:99–109.

Salmons, J. (2019). Laryngeal phonetics, phonology, assimilation and final neutralization. In Page, R. and Putnam, M. T., editors, *Cambridge Handbook of Germanic Linguistics*, pages 119–142. Cambridge University Press, Cambridge.

Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, 52:183–204.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the ICPhS*, pages 607–610, San Francisco.

Schilling-Estes, N. (2004). Constructing ethnicity in interaction. *Journal of Sociolinguistics*, 8:163–195.

Schleef, E. (2013). Glottal replacement of /t/ in two British capitals: Effects of word frequency and morphological compositionality. *Language Variation and Change*, 25:201–223.

Schoorman, H., Heeringa, W., and Peters, J. (2015). Regional variation of Saterland Frisian vowels. In *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Schultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *Journal of the Acoustical Society of America*, 132:EL95–101.

Seyfarth, S. and Garellek, M. (2018). Plosive voicing acoustics and voice quality in Yerevan Armenian. *Journal of Phonetics*, 71:425–450.

Shackleton, R. G. (2007). Phonetic variation in the traditional English dialects: A computational analysis. *Journal of English Linguistics*, 35:30–102.

Shimizu, K. (1996). *A cross-language study of the voicing contrasts of stop consonants in Asian languages*. Seibido, Tokyo.

Smith, J. and Holmes-Elliott, S. (2018). The unstoppable glottal: tracking rapid change in an iconic British variable. *English Language and Linguistics*, 22:323–355.

Solanki, V. J. (2017). *Brains in dialogue: investigating accommodation in live conversational speech for both speech and EEG data*. PhD thesis, University of Glasgow.

Solé, M.-J. (2007). Controlled and mechanical properties in speech. In Beddor, P. and Ohala, M., editors, *Experimental Approaches to Phonology*, pages 302–321. Oxford University Press, Oxford.

Sonderegger, M., Bane, M., and Graff, P. (2017). The medium-term dynamics of accents on reality television. *Language*, 93:598–640.

Sonderegger, M., Stuart-Smith, J., Knowles, T., MacDonald, R., and Rathcke, T. (2020a). Structured heterogeneity in Scottish stops over the twentieth century. *Language*, 96:94–125.

Sonderegger, M., Stuart-Smith, J., McAuliffe, M., Macdonald, R., and Kendall, T. (2020b). Managing data for integrated speech corpus analysis in SPeech Across Dialects of English (SPADE). In *Open Handbook of Linguistic Data Management*. MIT Press, Cambridge.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv e-prints*, page arXiv:1703.05339.

Stevens, M. and Harrington, J. (2014). The individual and the actuation of sound change. *Loquens*, 1:1–10.

Stuart-Smith, J. (2008). Scottish English: phonology. In Kortmann, B. and Upton, C., editors, *Varieties of English: British Isles*, pages 48–70. Mouton de Gruyter, Berlin.

Stuart-Smith, J., Jose, B., Rathcke, T., MacDonald, R., and Lawson, E. (2017). Changing sounds in a changing city: An acoustic phonetic investigation of real-time change over a century of Glaswegian. In Montgomery, C. and Moore, E., editors, *Language and a Sense of Place: Studies in Language and Region*, pages 38–65. Cambridge University Press, Cambridge.

Stuart-Smith, J., Sonderegger, M., Macdonald, R., Mielke, J., McAuliffe, M., and Thomas, E. (2019). Large-scale acoustic analysis of dialectal and social factors in English /s/-retraction. In *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia.

Stuart-Smith, J., Sonderegger, M., Rathcke, T., and Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, 6:505–549.

Summers, W. V. (1987). Effects of stress and final consonant voicing on vowel production: articulatory and acoustic analyses. *Journal of the Acoustical Society of America*, 82:847–863.

Sun, D. X. and Deng, L. (1995). Analysis of acoustic-phonetic variations in fluent speech using TIMIT. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Detroit, MI.

Swan, J. T. (2016). Canadian English in the Pacific Northwest: A phonetic comparison of Vancouver, BC and Seattle, WA. In *Proceedings of the 2016 Annual Conference of the Canadian Linguistics Association*, Calgary. University of Calgary.

Sweet, H. (1880). *A handbook of phonetics*. MacMillan & Co., London.

Tagliamonte, S. and Temple, R. (2005). New perspectives on an ol variable: (t, d) in British English. *Language Variation and Change*, 17:281–302.

Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24:135–178.

Takada, M. (2011). *Nihongo no gotou heisa'on no kenkyuu: VOT no kyoujiteki bunpu to tsuujiteki henka [Research on the word-initial stops of Japanese: synchronic distribution and diachronic change in VOT]*. Kurosio, Tokyo.

Takada, M., Kong, E. J., Yoneyama, K., and Beckman, M. E. (2015). Loss of prevoicing in Modern Japanese /g, d, b/. In *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2019a). Structured speaker variability in spontaneous Japanese stop contrast production. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Tanner, J., Sonderegger, M., Stuart-Smith, J., and Fruehwald, J. (2020). Towards "English" phonetics: variability in the pre-consonantal voicing effect across English dialects and speakers. *Frontiers in Artificial Intelligence*, 3.

Tanner, J., Sonderegger, M., Stuart-Smith, J., and SPADE-Consortium (2019b). Vowel duration and the voicing effect across English dialects. *University of Toronto Working Papers in Linguistics*, 41:1–13.

Tanner, J., Sonderegger, M., and Wagner, M. (2017). Production planning and coronal stop deletion in spontaneous speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 8.

Tauberer, J. and Evanini, K. (2009). Intrinsic vowel duration and the post-vocalic voicing effect: some evidence from dialects of North American English. In *Proceedings of Interspeech*.

Temple, R. (2014). Where and what is (t,d)? a case study in taking a step back in order to advance sociophonetics. In Celata, C. and Calami, S., editors, *Advances in Sociophonetics*, pages 97–136. John Benjamins, Amsterdam.

Theodore, R. M., Miller, J. L., and DeSteno, D. (2009). Individual talker differences in voice-onset-time: contextual influences. *Journal of the Acoustical Society of America*, 126:3974–3982.

Thomas, C. K. (1947). *An Introduction to the Phonetics of American English*. Ronald Press Company, New York.

Thomas, E. R. (2001). *An acoustic analysis of vowel variation in New World English*. American Dialect Society.

Thomas, E. R. (2003). Secrets revealed by Southern vowel shifting. *American Speech*, 78:150–170.

Thomas, E. R. (2006). Evidence from Ohio on the evolution of /æ/. In Murray, T. E. and Simon, B. L., editors, *Language Variation and Change in the American Midland: a new look at "Heartland" English*, pages 69–89. John Benjamins, Amsterdam.

Thomas, E. R. (2011). *Sociophonetics: An Introduction*. Palgrave MacMillan, Basingstoke.

Thomas, E. R. (2018). Acoustic-phonetic dialectology. In Boberg, C., Nerbonne, J., and Watt, D., editors, *Handbook of Dialectology*, pages 314–329. John Wiley and Sons.

Torreira, F. and Ernestus, M. (2011). Vowel elision in casual French: The case of vowel /e/ in the word c'était. *Journal of Phonetics*, 39:50–58.

Trude, A. M. and Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, 27:979–1001.

Trudgill, P. (1999). *The Dialects of England*. Blackwell, Oxford.

Tsujimura, N. (2014). *Introduction to Japanese Linguistics*. Wiley-Blackwell, Oxford.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America*, 58:434–445.

Vasishth, S., Mertzen, D., Jäger, L. A., and Gelman, A. (2018a). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

Vasishth, S., Nicenboim, B., Beckman, M., Li, F., and Kong, E. J. (2018b). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71:147–161.

Venditti, J. (2005). The J_ToBI model of Japanese intonation. In Sun-Ah, J., editor, *Prosodic Typology*, pages 172–200. Oxford University Press, Oxford.

Watson, C. I. and Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106:458–468.

Weinreich, U. (1954). Is a structural dialectology possible? *Word*, 10:388–400.

Weinreich, U., Labov, W., and Herzog, M. I. (1968). Empirical foundations for a theory of language change. In Lehmann, W. P. and Malkiel, Y., editors, *Directions for Historical Linguistics*, pages 95–195. University of Texas Press, Austin.

Wells, J. C. (1982). *Accents of English*. Cambridge University Press, New York.

Wetzell, B. (2000). Rhythm, dialects, and the Southern Drawl. Master's thesis, North Carolina State University.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91:1707–1717.

Williams, D., Elvin, J., Escudero, P., and Gafos, A. (2019). Multidimensional variation in English diphthongs. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia.

Williams, D. and Escudero, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *Journal of the Acoustical Society of America*, 136:2751–2761.

Winkelmann, R., Harrington, J., and Jğnsch, K. (2017). Emu-sdms: Advanced speech database management and analysis in r. *Computer Speech & Language*, 45(Supplement C):392 – 410.

Wolfram, W. and Beckett, D. (2000). The role of the individual and group in earlier African American English. *American Speech*, 75:3–33.

Wolfram, W. A. (1969). *A sociolinguistic description of Detroit Negro speech*. Center for Applied Linguistics, Washington DC.

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. *UC Berkeley Phonology Lab Annual Report*, pages 29–43.

Yu, A. and Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Reviews of Linguistics*, 5:131–150.

Yuan, J. and Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, 65:67–74.

Yuan, J., Liberman, M., and Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *Proceedings of Interspeech 2006*.

Yuan, J., Liberman, M., and Cieri, C. (2007). Towards an integrated understanding of speech overlaps in conversation. In *Proceedings of the International Congress of Phonetic Sciences XVI*, pages 1337–1340.

Zahorian, S. A. and Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94:1966–1982.

Zimmerman, S. A. and Sapon, S. M. (1958). Note on vowel duration seen cross-linguistically. *Journal of the Acoustical Society of America*, 30:152–153.