THE UNIVERSITY OF CHICAGO

PHONETIC AND PHONOLOGICAL DYNAMICS ON REALITY TELEVISION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
DEPARTMENT OF COMPUTER SCIENCE
AND
THE FACULTY OF THE DIVISION OF THE HUMANITIES
DEPARTMENT OF LINGUISTICS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
MORGAN SONDEREGGER

CHICAGO, ILLINOIS
AUGUST 2012

In memory of Partha Niyogi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My biggest thanks go to my two advisors during graduate school. Partha Niyogi's influence on how I think and do research has been profound. His intellectual rigor, knack for drawing surprising connections, deep knowledge, and kindness were inspirational, and I am among the many who will always be influenced by what he taught in his too-short life. When Partha was no longer able to serve as my advisor, I was fortunate that John Goldsmith stepped in to fill that role. Any student looking to straddle linguistics and computer science couldn't ask for a better advisor than John, who supervised this thesis. I have benefited greatly from his wisdom, unique perspective, and intellectual scope. I thank my committee members for all their help and advice on this thesis, and throughout grad school: Joseph Keshet, Jason Riggle, Anne Rogers, and Alan Yu.

I have received research and travel funding from a U.S. Department of Education GAANN Fellowship; grants from the NSF and Accenture to Partha Niyogi; a Daniels Fellowship and a McCormick Fellowship from the University of Chicago; and grants from the Department of Computer Science. I also acknowledge travel funding from the U. Chicago Office of Graduate Student Affairs, the Association for Computational Linguistics, Toyota Technological Institute-Chicago, the International Speech Communication Association, the Chicago Linguistic Society, and an NSF grant to Alan Yu.

I thank the international teams of volunteers behind the open-source programs `R` and Python, which were used extensively for this thesis. Many of my statistics questions were answered in the `r-sig-mixed-models` and `r-lang` lists, especially posts by Ben Bolker, Doug Bates, Roger Levy, and Florian Jaeger; others were answered by Rina Foygel, through the U. Chicago Department of Statistics consulting service.

The Big Brother chapters report work which is part of a larger collaboration with Max Bane and Peter Graff, and the original idea of examining linguistic dynamics on Big Brother is equally due to each of us. Without Max's and Peter's dedication, skills, and enthusiasm, the project would still be just an intriguing idea. Channel 4/Endemol provided permission to acquire the Big Brother data. Undergraduate research assistants have been crucial team members: Rachel Hwang, Adam Large, and especially Maria Nelson and Natalie Rothfels, who likely know more about this season of Big Brother UK than anyone in North America. Students in Jason Riggle's 2011 Phonological Analysis II class helped to annotate the coronal stop deletion data: Andrea Beltrama, Tasos Chatzikonstantinou, Alastair Cleve, Erin Franklin, Brett Kirken, Jackson Lee, Maria Nelson, Krista Nicoletto, Talia Penslar, Hannah Provenza, and Natalie Rothfels. Thanks, everyone.

Nattalia Paterson, Matt Goldrick, Hugo van Hamme, and Chi-Yueh Lin generously provided data used in the experiments in Chapter 3.

I have benefited from discussions and generosity from too many other colleagues to list, including Max Bane, James Kirby, Peter Graff, Ross Girshick, Ankan Saha, Adam Albright, Klinton Bicknell, Paula Chesley, Matt Carlson, Robert Daland, Matt Goldrick, Kyle Gorman, John Hale, Jeff Heinz, Florian Jaeger, Daniel Ezra Johnson, Shira Katseff, Jon Keane, Christina Kim, Tyler Kendall, Chris Kennedy, Karen Livescu, Ben Munson, Janet Pierrehumbert, Sravana Reddy, Tyler Schnoebelen, Charlotte Vaughn, Colin Wilson, and Kristine Yu.

Outside academia, thanks to my friends for supporting me, especially Laura Hodges and Alix Midgley. Finally, thanks to my parents for their support, and for most of the qualities which got me this far.

ABSTRACT

The sounds of a language spoken by an individual, or shared by a speech community, can be seen as both remarkably stable and subject to great change. For example, one's accent intuitively seems very stable over adulthood, especially in comparison to one's constantly changing vocabulary; however, many people report that their accents shifted after moving from one city to another, or due to social pressure. This thesis addresses two questions about stability and change in sound systems, or *phonetic and phonological dynamics*: what are the dynamics of sound systems in individuals during adulthood, and what causes underly them?

Previous work has addressed these questions on two timescales. Short-term studies examine shifts in (phonetic and phonological) variables under exposure to the speech of others, for example over the course of a conversation. Long-term studies examine shifts in variables between time points separated by years. Short-term shifts are fairly robust, with most speakers showing some shift for most variables. Long-term shifts are extremely irregular, with huge variation in the amount of shift among speakers and variables. What is the relationship between the different patterns seen in short-term and long-term dynamics? And more generally, what do the dynamics of sound systems look like at any time scale in between?

The bulk of the thesis is a "medium-term" case study addressing these questions, using a setting where day-by-day phonetics and phonological dynamics can be observed within individuals: the reality television show Big Brother UK, where speakers live in an isolated house for three months, and are continuously recorded. We consider five variables in six hours of speech from one season of the show: voice onset time, coronal stop deletion, and formant frequencies for three vowels. We build mixed-effect regression models of day-to-day time dependence for each variable, for each of 12 speakers, controlling for linguistic factors. Variability is the norm: speakers and variables show qualitatively different types of time dependence, with a significant minority showing stability. There is some evidence that particular speakers (across variables) and particular variables (across speakers) show characteristic types of time dependence, and that some time dependence is due to style shifting. Long-term time trends do sometimes occur, which could be due to accumulation of short-term shifts. Day-by-day variation is common, but far from universal. These results suggest a tentative account of the relationship between short-term and long-term dynamics, and directions for future work.

The thesis also addresses two topics closely related to phonetic and phonological dynamics: synchronic variation, and automatic phonetic measurement. For each variable, we build a model of synchronic variation as a preliminary step to modeling the variable's dynamics within speakers; the models for voice onset time and coronal stop deletion turn out to yield interesting and surprising findings with respect to previous work. Many questions of interest about phonetic and phonological dynamics require radically scaling up from the hand-labeled datasets used in most previous work, making automatic measurement methods crucial. Our main methodological contribution is a discriminative, large-margin algorithm for automatic VOT measurement, treated as a case of predicting structured output from speech. The algorithm is tested on data from four corpora representing different types of speech. It achieves performance near human intertranscriber reliability, and compares favorably with previous work.

CHAPTER 1

INTRODUCTION

The sounds of a language spoken by an individual, or shared by a speech community, can be seen as both remarkably stable and subject to great change. The vowels of English have been constantly changing since the Middle Ages, and vary between dialects, yet the consonants of English have remained largely the same for several centuries. At the level of an individual, one's accent intuitively seems very stable over adulthood, especially in comparison to one's constantly changing vocabulary; however, many people report that their accents shifted after moving from one city to another, or due to social pressure, or for no obvious reason. These are all cases of stability and change in sound systems, or phonetic and phonological dynamics. At a high level, this thesis addresses two questions: what are the dynamics of sound systems, and what causes underlie them?

These questions are closely linked to the study of synchronic phonetic and phonological variation. There is tremendous variation in how sounds are realized at a given time, as a function of *static factors*: linguistic context, social factors, processing factors, speaking style, and so on. Phonetic and phonological dynamics constitute variation as a function of time, or *longitudinal variation*. However, any pattern of variation observed over time is a function of both time and static factors. To determine the underlying time dependence, the effects of static factors must be determined, and controlled for. The entanglement of time and static factors is especially acute in corpus data, where many static factors vary at once. There is also a more fundamental connection between synchronic variation and dynamics: all linguistic change begins in patterns of synchronic variation (though not not all synchronic patterns lead to change). Thus, understanding the causes of an observed pattern of longitudinal variation, entails understanding its connection to corresponding patterns of synchronic variation.

The questions which can be asked about phonetic and phonological dynamics, particularly in corpora, fundamentally depend on the methods used for measurement and statistical analysis. Large datasets are already needed in corpus studies of synchronic variation, to deal with the simultaneous presence of many factors; much larger datasets are needed to search for time dependence, because the system must be characterized at multiple points in time. The resulting datasets are both large and complex. To scale up to the amount of data needed, automatic measurement methods are crucial. Simultaneously characterizing the effects of time and static factors requires powerful statistical methods.

1.1 Summary

The bulk of this thesis is a case study of phonetic and phonological dynamics in individuals over three months. The setting is a 'natural experiment': the reality television show Big Brother, where individuals live in an isolated house and are continuously recorded. The show provides an excellent setting for studying the dynamics of sound systems and their sources.

This thesis consists of three conceptual pieces, corresponding to the division above: methodology, synchronic variation, and longitudinal variation. The pieces overlap somewhat, for ease of presentation. The first gives background on statistical methods, and presents a new algorithm for automatic phonetic measurement. The second and third examine synchronic variation and longitudinal variation in a corpus of spontaneous speech from the house, in five variables: voice onset time (VOT), coronal stop deletion (CSD), and formant frequencies for three vowels.

1.1.1 Methodology

Chapter 2 provides background on statistical methods used in this thesis, in particular mixed-effects regression models. Mixed models generalize classical regression for the case of data with an underlying grouping structure. When modeling phonetic variation, accounting for the grouping structure of the data (e.g., by word, speaker, or both) is crucial for accurately determining both time dependence and the effects of static predictors.

Chapter 3 presents our main methodological contribution: a machine learning algorithm for automatic VOT measurement, treated as a case of predicting structured output from speech. Manually-labeled data is used to train a function that takes as input a speech segment of an arbitrary length containing a voiceless stop, and outputs its VOT. The function is explicitly trained to minimize the difference between predicted and manually-measured VOT; it operates on a set of acoustic feature functions designed based on spectral and temporal cues used by human annotators. The algorithm is applied to initial voiceless stops from four corpora, representing different types of speech. Using several evaluation methods, the algorithm's performance is near human intertranscriber reliability, and compares favorably with previous work. Furthermore, the algorithm's performance is minimally affected by training and testing on different corpora, and remains essentially constant as the amount of training data is reduced to 50–250 manually-labeled examples, demonstrating the method's practical applicability to new datasets.

Variation in VOT and vowel formants in the Big Brother corpus are primarily studied using automatic measurements. Automatic formant measurements are made following forced alignment of each sound file to a sequence of phonemes corresponding to its orthographic transcription; both steps use modified versions of existing software.

1.1.2 Synchronic variation

Chapters 4 and 5 contain preliminaries to modeling phonetic variation in Big Brother. Chapter 4 reviews relevant areas of previous work:

- *Previous linguistic studies using reality TV data* have mostly studied pragmatic or discourse phenomena related to the performative aspect of reality television. Our study adds to a smaller strand where a show is used as a natural experiment, to address questions well-aligned with the show's structure.

- *Phonetic and phonological variation in spontaneous speech* has been studied in two largely distinct literatures, depending on whether the primary interest is in phonetics and phonological phenomena, or variation and change in speech communities.[1] Our study falls into both categories.

- *Longitudinal phonetic and phonological variation during adulthood* is discussed below.

- *Factors conditioning variation* in each of the five variables have been extensively studied in the phonetics and sociolinguistics literatures.

Chapter 5 describes the structure of the show, our corpus of spontaneous speech from the show, and datasets of phonetic measurements for each of the five variables.

Chapter 6 presents models of synchronic variation in each variable, as a function of static factors related to the host word, the speaker, surrounding segments, or speaking style. We also model arbitrary differences between speakers and words, beyond the effects of static factors. As described above, a first motivation for building these *static models* is as a preliminary step to building *dynamic models* of longitudinal variation, to determine how to control for static factors.

A second motivation is that the static models are interesting in their own right. The models for voice onset time and coronal stop deletion in particular yield novel and surprising findings with respect to previous work. Both have largely been studied in North American varieties of English, while our dataset contains an unusual mix of accents, largely from non-standard British dialects. VOT has been examined almost exclusively in laboratory studies of planned speech, where the effects of a small set of factors are precisely determined, with all others held constant. By examining VOT in a spontaneous speech corpus, we are able to determine the relative importance of different conditioning factors detailed by laboratory studies, and how much these factors vary by speaker. CSD has usually been studied in spontaneous speech, but for speakers from a single speech

[1] Of course, these categories are not mutually exclusive.

community. Our unusual dataset compels us to build more complex models than used in previous studies, allowing both base deletion rate and the effects of conditioning factors to vary between speakers. The models make interesting predictions for several conditioning factors, and our results raise a number of methodological points with relevance for the interpretation of previous work on CSD, and for the design of future studies of CSD.

1.1.3 Longitudinal variation

The heart of this thesis is Sec. 4.3 and Chapter 7, which address the two questions raised at the outset about longitudinal variation, specifically within adults (as are all speakers in the Big Brother house). To situate this study with respect to previous work, it is convenient to rephrase these two questions as three more specific questions:

1. What time dependence do different phonetic and phonological variables show within individual speakers?

2. How and why do individuals differ in the dynamics of particular variables?

3. How and why do variables differ in their dynamics within individuals?

Sec. 4.3 reviews previous work addressing these questions, which can be divided into two categories corresponding to different timescales. *Short-term* studies examine shifts in phonetic and phonological variables in one's speech under exposure to the speech of others, over the course of a conversation or a laboratory experiment. An important motivation for many short-term studies is the hypothesis that sound change, both in individuals and in communities, results in part from an accumulation of the short-term shifts which occur in interaction. This account rests on the *persistence hypothesis*, that such shifts can and do accumulate within individuals. *Long-term* studies examine shifts in phonetic and phonological variables between several time points separated by years (potentially many), for individuals who either remain in the same speech community or move between dialect regions. The interpretation of long-term studies rests on the *stationarity* *hypothesis*, that individuals do not fluctuate much in their baseline use of a variable from day to day, so any change observed after controlling for conditioning factors is meaningful. However, very little previous work tests either hypothesis, and it is not currently known under what conditions either one holds.

Short-term shifts are fairly robust: for most phonetic and phonological variables, most speakers show some shift. Long-term shifts are extremely irregular: there is huge variation in the amount of shift shown by different individuals for different variables. In many cases a majority show stability, with a minority showing significant change. Yet given the robustness of short-term shifts, if the hypothesis that such shifts accumulate holds, long-term shifts should be the norm. How can the patterns seen in short-term and long-term change be reconciled, and where does the disconnect lie?

Big Brother offers a case of *medium-term* change, where the dynamics of phonetic and phonological variables can be observed within individuals over three months, and used both to test the persistence and stationarity hypotheses, and to address the mismatch between short-term and long-term change. Chapter 7 models day-to-day time trajectories for our five variables for 12 speakers. For each speaker/variable pair, a best *dynamic model* of time dependence, controlling for static factors, is determined. We find that variability is the norm: speakers and variables show four qualitatively different types of time dependence, with a significant minority showing stability. Thus, the dynamics look more like long-term change than short-term change. There is some evidence that particular speakers (across variables) and particular variables (across speakers) show characteristic types of time dependence. These results support weaker forms of the persistence and stationarity hypotheses: long-term time trends do sometimes occur, which could be due to accumulation of short-term shifts. Day-by-day variation is common, but far from universal. We give a tentative account of the relationship between short-term and long-term dynamics in individuals, which addresses the apparent mismatch between them, and suggests directions for future work.

1.2 Terminology: Phonetics and phonology

This thesis examines variation in *sound systems*, by which we mean both phonetics and phonology, for either an idiolect or a variety spoken in a speech community. Traditionally, the domain of phonology is the organization of sounds at an abstract level (phonemes) in a language, and the domain of phonetics is the physical realization of sounds (phones). However, the boundary between the two has become very murky. Though most scholars would still acknowledge that a distinction exists (some phenomena belong squarely in phonetics or phonology), "the field has reached no consensus about what the [phonetics-phonology] interface is, nor is it even agreed that one exists at all" (Kingston, 2007: 401).[2]

The lack of consensus is particularly acute for variation, especially across the multiple subfields addressed by this thesis (phonetics, sociolinguistics, phonology). "Phonetic variation" traditionally encompassed variability in the surface realization of a phoneme, but much "phonological variation" studied in sociolinguistics since the 1960s falls into this category. "Phonological variation" has been increasingly studied by phonologists as well (Coetzee and Pater, 2011), but there is no consensus on what the term encompasses, e.g. only categorical differences or gradient subphonemic variation as well.[3]

Fortunately, the distinction between phonetics and phonology is largely irrelevant for the studies in this thesis. Unfortunately, there is no established term for 'phonetic and/or phonological', and this adjective is awkward to use throughout. Rather than invent a new one, we will tend to use 'phonetic' (i.e., 'phonetic variable', 'phonetic variation'), simply because we will mostly investigate variation in each of our five variables in isolation, and not as part of a system of units. Thus, traditionally phonological notions such as contrast and alternations do not come into play, while many traditionally phonetic notions

[2] The second clause encompasses two extreme positions: that there is total overlap between phonetics and phonology (Ohala, 1990), or no overlap at all (Hale and Reiss, 2000).

[3] Coetzee and Pater's review in fact defines it as "a situation in which a single morpheme can be realized in more than one phonetic form in a given environment" (401), which would subsume any definition of phonetic variation.

do. However, this shorthand should not be taken to imply any assumptions about the extent to which any of our variables are phonetic or phonological in any meaningful sense. When discussing variation in sound systems more generally, or making points based on our variables with broader implications for such variation, we will tend to use ‘phonetic and phonological’.

CHAPTER 2

STATISTICAL BACKGROUND

This chapter briefly introduces statistical methods used in this thesis, with particular attention to mixed-effects regression models, which we will use extensively.[1]

2.1 Preliminaries

Linear regression In classical linear regression, we are given n observations. The i^{th} observation consists of a real-valued *response*, y_i, and values for p *predictors*, $(x_i^0, x_i^1, \ldots, x_i^{p-1})$. It is typically assumed that the first predictor, the *intercept*, is always 1 ($x_i^0 = 1$). The *model matrix* X is the $n \times p$ matrix whose $(i, j+1)$ entry is x_i^j. The response is modeled as a linear function of the predictors, plus an error term:

$$y_i = \beta^0 + \beta^1 x_i^1 + \cdots + \beta^{p-1} x_i^{p-1} + \epsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta^0, \beta^1, \ldots, \beta^{p-1})$ is the $p \times 1$ vector of *regression coefficients*, and the ϵ_i are i.i.d. drawn from $N(0, \sigma^2)$. An estimate of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}$, is determined, often using ordinary least squares. The resulting *fitted values* are $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$, and the *residuals* are $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

Logistic regression In classical logistic regression, a categorical response is modeled as a function of a set of predictors. Following common practice, we use “logistic regression” as shorthand for “binomial logistic regression”, where the response is binary. For each observation, the response $y_i \in \{0, 1\}$ is assumed to follow a Bernoulli distribution whose expectation is an unseen parameter, p_i. A function of p_i, the *link function*, is assumed to be

[1] For more comprehensive treatments, see e.g., Gelman and Hill (2007); Maindonald and Braun (2007) for statistical methods and Gelman and Hill (2007); Snijders and Bosker (2011) for mixed-effects models. From the perspective of analyzing linguistic data, see e.g., Baayen (2008); Johnson (2011); Levy (2012).

a linear function of the predictors. A common choice for binomial data is the *logistic link*:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta^0 + \beta^1 x_i^1 + \cdots + \beta^{p-1} x_i^{p-1}, \quad i = 1, \ldots, n \tag{2.2}$$

Note that there is no error term, in contrast to Eq. (2.1). The left-hand side is often written logit (p_i), or called the log-odds of p_i. The logit function has range $(-\infty, \infty)$.

As for linear regression, an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is determined. Given $\hat{\boldsymbol{\beta}}$ and a vector $\boldsymbol{x}$ of predictor values, $\hat{\boldsymbol{\beta}} \cdot \boldsymbol{x}$ is the predicted log-odds that the response is 1. Thus, the model's prediction (which can be any real number) is not directly comparable to the response (which is 0 or 1), unlike in linear regression.

Types of variables Following Gelman and Hill (2007: 37), we distinguish between inputs and predictors. *Inputs*, or *input variables*, are "information on the units that go into the predictors", which we denote using using SMALL CAPS. Predictors are functions of the inputs, which make up the columns of the model matrix.

For example, suppose VOT for word-initial voiceless stops is modeled as a function of three inputs: speaking rate (RATE), word frequency (FREQUENCY), and the stop's place of articulation (POA). RATE and FREQUENCY are continuous variables, while POA is a categorical input, or *factor*, which can take on a number of possible values, called *levels*. We denote factor levels in `teletype`. (POA has 3 levels: `bilabial`, `alveolar`, `velar`.)

To include a factor x with k levels in a regression model, it is coded as $k-1$ numerical predictors, or *contrasts*.[2] The number of contrasts is $k-1$ because the intercept is a predictor as well, so which of the k levels of x a observation takes on is determined by k predictors. Thus, a factor with k levels corresponds to $k-1$ degrees of freedom. The choice of a contrast coding scheme defines the interpretations of the $k-1$ predictors, in terms of the value which the response takes on for different factor levels.

[2] For more on contrast coding, see e.g., Cohen et al. (2003: Ch. 8–9); Venables and Ripley (2002: Sec. 6.2).

Two types of contrast coding are used in this thesis. In *dummy coding*, one level is chosen as the base, and the contrasts represent the difference between each other level and the base. For example, if `bilabial` is chosen as the base for the dummy-coded contrasts for POA, the interpretation of the two contrasts is "difference between `alveolar` and `bilabial`" and "difference between `velar` and `bilabial`." In *Helmert coding*, given an ordering of the factor's levels, each contrast represents the difference between a level and the mean of all previous levels. For POA, the interpretation of the two contrasts is "difference between `alveolar` and `bilabial`" and "difference between `velar` and the mean of `bilabial` and `alveolar`." Importantly, different choices of contrast for a factor included in a regression are simply reparametrizations of the same statistical model. The interpretation of the regression coefficients for the contrasts and the intercept change, but the predictions of the model do not.

Interactions It is often of interest to model the effect of one input on the response as conditional on the value of another input. For example, the effect of speaking rate on VOT might depend on the stop's place of articulation. This is accomplished by adding predictors for the *interaction* of the two inputs. When predictions for interactions are included in a regression, the non-interaction predictors are called *main effects*. For two continuous inputs, the interaction is simply the product of the two main effects, called a two-way interaction; for three continuous inputs, it is the product of the three main effects (three-way interaction), etc. More generally, for two inputs corresponding to k and j main effects (i.e., number of degrees of freedom), their interaction corresponds to kj predictors.

To give a concrete example, consider a classical linear regression model of VOT as a function of RATE, POA (dummy-coded, as described above), and their interaction, for n words. y_i is the VOT of word i, and there are six predictors (Table 2.1).

Table 2.1: Predictors for RATE, POA, and their interaction for observation i.

Predictor	Interpretation	Note
x_i^0	1 (the intercept)	
x_i^1	Speaking rate for word i	
x_i^2	POA contrast 1	1 if POA=`alveolar`, 0 otherwise
x_i^3	POA contrast 2	1 if POA=tttvelar, 0 otherwise
x_i^4	$x_i^1 \cdot x_i^2$	
x_i^5	$x_i^1 \cdot x_i^3$	

2.2 Mixed-effects regression

Mixed-effects regression models, or mixed models (also known as hierarchical, or multilevel models), are a generalization of classical regression to the case where observations have an underlying grouping structure, possibly hierarchical. For example, for most models in this thesis, each observation represents a phonetic variable measured for a word spoken by a particular speaker, so each point is grouped by WORD and SPEAKER. These are *grouping levels*, which consist of a set of *units* (i.e., individual speakers). Intuitively, each observation's response is modeled both as a function of properties of the observation, and of properties of the group units it belongs to.

Mixed-effects models contain two types of coefficients: *fixed effects* do not vary between the units in a group, while *random effects* do. Fixed-effect coefficients are analogous to the elements of the coefficient vector $\boldsymbol{\beta}$ in classical regression; they measure the mean effect of each predictor across units in a grouping level. Random-effect coefficients are assumed to be drawn from a probability distribution; they measure how much each unit deviates in the effect of some predictor from the overall mean.

2.2.1 Models

A general treatment of mixed-effects models, even simple ones, would require introducing an overabundance of notation for our purposes. Instead, we will give examples of the types of models used in this thesis, introducing terminology as we go.

2.2.1.1 Linear mixed-effects models: Single grouping level

Suppose we model n observations of VOT from word-initial stops in K words spoken by J speakers, as a function of three inputs: speaking rate (RATE), speaker gender (GENDER), and word frequency (FREQUENCY). Let y_i be the response and $s[i]$ and $w[i]$ denote the word and speaker of the i^{th} observation. Let x_i^1 be the RATE for observation i, $x_{s[i]}^2$ the GENDER of speaker $s[i]$ (dummy-coded: 1/0 for females/males), and $x_{w[i]}^3$ the FREQUENCY of word $w[i]$. Finally, let RATE and FREQUENCY be centered, so a value of 0 for x_i^1 means a word of average frequency spoken at an average rate. RATE is an *observation-level* input: an observed quantity which differs (potentially) between observations. Similarly, GENDER is a speaker-level input and FREQUENCY is a word-level input.

Model 1: Random intercept only In a first model, we ignore that observations are grouped by word. We assume that VOT varies as a function of RATE and FREQUENCY in the same way across all speakers, and as a function of GENDER in the same way across all words. In addition, we assume each speaker's characteristic VOT differs from the mean by an offset, $b_{s[i]}$. Each observation's VOT is modeled as:

$$\begin{aligned}
&y_i = \beta^0 + \beta^1 x_i^1 + \beta^2 x_{s[i]}^2 + \beta^3 x_{w[i]}^3 + b_{s[i]} + \epsilon_i, \quad i = 1, \ldots, N \\
&b_j \sim N(0, \sigma_0^2), \quad j = 1, \ldots, J \\
&\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, N
\end{aligned}$$

The coefficients β^1, β^2, and β^3 are fixed effects. They describe the *slopes* for speaking rate (how much VOT increases for an increase of 1 unit), word frequency (similar), and gender (the difference in VOT between male and female speakers).

The coefficients $b_1, \ldots, b_J$ are random effects, specifically *random intercepts*, in this case "by-speaker random intercepts." They quantify how much each speaker's VOT differs from the group-level mean, β^0. Thus, for a word of average frequency spoken at an average speaking rate, the VOT for male speaker j is $\beta^0 + b_j$. It is assumed that speakers' offsets are normally distributed. There is also an observation-level error term, also assumed to be normally distributed.

Assuming the speakers whose data were modeled were representative of a larger population, $\hat{\sigma_0}^2$ is the model's prediction of how much speakers in the population vary in their characteristic VOT values, after accounting for the effect of GENDER. For example, 95% of male speakers in the population are predicted to have VOTs in $(\hat{\beta}^0 - 2\hat{\sigma}_0, \hat{\beta}^0 + 2\hat{\sigma}_0)$.

Model 2: Random intercept and random slopes, uncorrelated In Model 1, it was assumed that speakers all had the same slope for speaking rate, β^1, but had varying intercepts. Now, assume that both the intercept and the slope for speaking rate vary by speaker, with $b^0_{s[i]}$ and $b^1_{s[i]}$ the offsets for speaker $s[i]$. The model is now:

$$
\begin{aligned}
&y_i = (\beta^0 + b^0_{s[i]}) + (\beta^1 + b^1_{s[i]})x^1_i + \beta^2 x^2_{s[i]} + \beta^3 x^3_{w[i]} + \epsilon_i, \quad i = 1, \ldots, n \\
&b^0_j \sim N(0, \sigma^2_0), \quad j = 1, \ldots, J \\
&b^1_j \sim N(0, \sigma^2_1), \quad j = 1, \ldots, J \\
&\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n
\end{aligned}
$$

This model contains two types of random effects. $b^0_1, \ldots, b^0_J$ are by-speaker random intercepts, as in Model 1. Analogously, $b^1_1, \ldots, b^1_J$ are by-speaker *random slopes*, which measure how much each speaker's effect of speaking rate on VOT differs from the group-level mean, β^1. The by-speaker random intercepts, by-speaker random slopes, and observation-level errors are all normally distributed, with different variances. Further, the two by-speaker random effects are assumed to be uncorrelated across speakers; for example, it is not the case that speakers with higher VOTs also have larger slopes for speaking rate.

Model 3: Random intercept and random slopes, correlated It may not be realistic to assume uncorrelated random effects, for example if a clear correlation between VOT and the effect of speaking rate is observed across speakers in the empirical data. If the random slopes and random intercepts are allowed to covary, the model becomes:

$$\begin{aligned}
y_i &= (\beta_0 + b^0_{s[i]}) + (\beta_1 + b^1_{s[i]})x^1_i + \beta_2 x^2_{s[i]} + \beta_3 x^3_{w[i]} + \epsilon_i, \quad i = 1, \ldots, n \\
\begin{pmatrix} b^0_j \\ b^1_j \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_0 & \sigma^2_{01} \\ \sigma^2_{01} & \sigma^2_1 \end{pmatrix} \right), \quad j = 1, \ldots J \\
\epsilon_i &\sim N(0, \sigma), \quad i = 1, \ldots, n
\end{aligned}$$

The two by-speaker random effects are now drawn from a multivariate normal distribution with mean **0**, and a 2×2 covariance matrix. Note that both Model 1 and Model 2 are special cases of Model 3, for different assumptions about this covariance matrix: $\sigma_1 = \sigma_{01} = 0$ gives Model 1, and $\sigma_{01} = 0$ gives Model 2.

Let Σ denote the covariance matrix of the by-speaker random effects. We could also choose to allow the slope for word frequency to vary by speaker, in which case Σ would be 3×3. Models 1–3 easily generalize to arbitrary many speaker-level and observation-level predictors. A by-speaker random slope can be added, or not, for each observation-level predictor.

2.2.1.2 Linear mixed-effects models: Multiple grouping levels

Most mixed-effects models in this thesis have two or more grouping levels. We give an example for the case of two levels; the extension to more is straightforward.

Model 4: Two grouping levels, random intercepts only Most mixed models in this thesis have two grouping-levels: speaker and word. In this setting, there can be both by-word and by-speaker random effects. Such models, where a unit in one grouping level can occur with multiple units in another grouping level, are said to have *crossed random effects* (Baayen et al., 2008; Hox, 2010). (For example, many words will be spoken by more than one speaker.)

Remaining with the VOT example, assume the same setup and notation as Model 1 (fixed effects for RATE, GENDER, FREQUENCY; by-speaker random intercepts), but now allow by-word random intercepts, c_k $(k = 1, \ldots, K)$. The model is now:

$$\begin{aligned}
y_i &= (\beta^0 + b_{s[i]} + c_{w[i]}) + \beta^1 x_i^1 + \beta^2 x_{s[i]}^2 + \beta^3 x_{w[i]}^3 + \epsilon_i, \quad i = 1, \ldots, N \\
b_j &\sim N(0, \sigma_0^2), \quad j = 1, \ldots, J \\
c_k &\sim N(0, \tau_0^2), \quad k = 1, \ldots, K \\
\epsilon_i &\sim N(0, \sigma^2), \quad i = 1, \ldots, N
\end{aligned}$$

In this model, both words and speakers are assumed to have characteristic VOT values; their offsets from the mean are normally distributed, with different variances. So for example, for a female speaker j and a word k with average frequency spoken at an average rate, VOT is $\beta^0 + \beta^2 + b_j + c_k$. $\hat{\tau}_0^2$ and $\hat{\sigma}_0^2$ are the model's estimates of how much words and speakers in the larger population vary in their characteristic VOT values, after accounting for word-level and speaker-level predictors.

It would be possible to add both by-speaker and by-word random slopes to Model 4, analogously to Models 2–3. In the current example, we could include a by-word random slope for GENDER, which would allow the VOT difference between men and women to vary by word. The models in this thesis do not include by-word random slopes, so no such examples are given here.

2.2.1.3 Mixed-effects logistic regression

Models 1–4 assume a linear response. As in the classical regression case, it is also possible to model a binomial response. Again, we now assume that the response $y_i \in \{0, 1\}$ is drawn from a Bernoulli distribution whose expectation is an unknown parameter, p_i. Assuming a logistic link, logit(p_i) is modeled almost exactly as y_i was in Models 1–4: as a linear function of the predictors, with coefficients consisting of fixed-effect terms, random effect terms, or a mixture of both. However, as in classical logistic regression, there is no error term (ϵ_i).

2.2.2 *Model fitting and results*

For an arbitrary linear mixed model, let $\boldsymbol{\beta}$ be the vector of fixed effect coefficients, and let σ^2 be the observation-level variance. Let $\boldsymbol{\Psi}$ be the vector consisting of all the random-effect variances, concatenated. (For example, for Model 4, $\boldsymbol{\Psi} = (\sigma_0^2, \tau_0^2)$.) A closed-form likelihood function can be written for the model in terms of these three parameters. Fitting the model results in parameter estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Psi}}$, and $\hat{\sigma}^2$, and corresponding standard errors. We fit mixed models using the `lme4` package in `R` (Bates et al., 2011), which chooses estimates which maximize either log-likelihood, or restricted maximum likelihood (Pinheiro and Bates, 2000: Sec 2.2). For mixed-effects logistic regression, there is no residual variance, so only estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ are produced. Because the likelihood function in this case does not have a closed-form solution, `lme4` optimizes a Laplace approximation to the likelihood (Bates, 2011).

BLUPs Although the random-effect variances $\boldsymbol{\Psi}$ are model parameters, the random effects themselves are not; they are unobserved parameters. However, it is often of interest to have estimates of the random effects themselves, for example to characterize how much individual speakers deviate from the norm. Estimates known as *best linear unbiased predictors* (BLUPs) are commonly used; these are the conditional modes of the random effects, conditional on the model parameters, evaluated at their estimated values ($\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Psi}}$, $\hat{\sigma}^2$) (Pinheiro and Bates, 2000: Sec. 2.2)

2.2.3 *Statistical significance*

We assume familiarity with the concepts of hypothesis tests, confidence intervals, and measures of statistical significance, for example in the context of classical linear regression. In summarizing models in this thesis, we will give measures of significance of sets of random-effect terms, individual fixed-effect coefficients, and sets of both random-effect terms and fixed-effect coefficients. In each case, various measures of significance have been proposed; we present a few which are both easy to calculate and widely used. In general, our impression is that there is no consensus yet on the best way to quantify and report uncertainty for mixed models.

Random effects The significance of a set of k random-effect terms (i.e., corresponding to k random effect variances) can be assessed by comparing the full model (where the terms are included) with a subset model (where the terms have been dropped). In the large-sample limit, twice the log of the ratio of the two models' likelihoods follows a χ^2 distribution with k degrees of freedom. Thus, in a *likelihood ratio test*, this quantity can be compared to a $\chi^2(k)$ distribution to obtain a p-value (Pinheiro and Bates, 2000: Sec. 2.4).

Fixed effects Let $\hat{\beta}$ and SE($\hat{\beta}$) be the parameter estimate and standard error for some fixed-effect coefficient in a mixed model. Under conditions which are usually satisfied in practice, $\hat{\beta}$/SE($\hat{\beta}$) follows a standard normal distribution (Pinheiro and Bates, 2000: Sec. 2.3). Thus, a (two-sided) Wald test can be applied to give a measure of significance.

Another more complex possibility does not rely on large sample size. Assuming some prior distribution over the model parameters ($\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, and $\hat{\sigma}^2$), it is possible to define a posterior distribution over these parameters conditional on the data. The `mcmcsamp` function in `lme4` performs Markov Chain Monte Carlo (MCMC) sampling from this posterior distribution, assuming locally-flat or locally non-informative priors over all model parameters. Sampling from the posterior gives N times gives N draws of $\hat{\beta}$. One minus the percentage of draws which are on the same side of zero as the fixed-effect estimate of $\hat{\beta}$ gives an *MCMC p-value*, which intuitively measures the probability that we are mistaken about the sign of $\hat{\beta}$.

Of these two methods, MCMC p-values better characterize uncertainty in the estimated value of a fixed effect. However, at this writing sampling from the posterior has only been implemented for linear mixed models with uncorrelated random effects.

The significance of a set of j fixed-effect terms can be assessed by a likelihood ratio test, similarly to a set of random-effect terms. In the large sample limit, the log of the likelihood ratio of a model with and without the j terms follows a $\chi^2(j)$ distribution.

Both fixed and random effects Finally, it is sometimes of interest to assess the significance of j fixed-effect terms and k random-effect terms. Let L_{k+j}, L_k, and L_0 be the likelihoods of the full model (including both types of terms) M_{k+j}, the model M_k where the fixed-effects terms have been dropped, and the model M_0 where both types of terms have been dropped. The difference in log-likelihood between the full model and the smallest model is

$$\log(L_{k+j}) - \log(L_0) = \underbrace{(\log(L_{k+j}) - \log(L_k))}_{\sim\chi^2(j)} + \underbrace{(\log(L_k) - \log(L_0))}_{\sim\chi^2(k)}$$

where the terms on the right-hand side follow χ^2 distributions because they correspond to dropping j fixed-effect terms and k random-effect terms, respectively. The sum of two independent random variables which follow $\chi^2(k)$ and $\chi^2(j)$ distributions follows a $\chi^2(j + k)$ distribution. Thus, assuming that the contribution of the fixed-effect and random-effect terms to the model are independent, the difference in log-likelihood between M_0 and M_{k+j} can be compared to a $\chi^2(j + k)$ distribution to obtain a p-value.

2.2.4 *Goodness of fit*

It is often of interest to measure the *goodness of fit* of a statistical model to the data. The most familiar measure is R^2, which applies to (classical) linear regression models. R^2 can be defined in several ways, all of which are equivalent for linear regression. The most common definition is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})} \tag{2.3}$$

where $\hat{y}_i$ is the model's prediction for observation i, and $\bar{y}$ is the mean over the y_i. Intuitively, the denominator in the fraction is the amount of variability in the data, after centering (i.e., the amount of variability remaining if the only predictor is the intercept), and the numerator is the amount of variability in the data which is unaccounted for by the model. Under this definition R^2 is the "fraction of variation explained" by the model.

Another definition uses the relative likelihood of the intercept-only and full models:

$$R^2 = 1 - (\frac{L_0}{L_M})^{\frac{2}{n}} \tag{2.4}$$

where L_0 and L_m are the likelihoods of these two models (Magee, 1990). Under this definition, R^2 is the amount of improvement of the full model over a baseline model, measured by their likelihood ratio.

Under either definition, R^2 ranges between 0 (no improvement over intercept-only model) and 1 (perfect fit to the data).

For logistic regression, there is no equivalent of R^2: a single quantity which ranges between 0 and 1, can be interpreted as percentage variance explained or improvement over a baseline model, and so on. A number of "pseudo-R^2" measures have been proposed, corresponding to different definitions of R^2 in the linear regression case. One widely-used option (Cox-Snell R^2) simply uses Eq. (2.4), since L_0 and L_M are also defined for logistic regression. However, the likelihood function for logistic regression models is such that the maximum value of L_M is 1; thus, the Cox-Snell R^2 has a maximum value less than 1.

Nagelkerke (1991) proposed rescaling R^2 by its maximum value, so the resulting measure lies between 0 and 1 (as for linear regression):

$$R^2_\nu = \frac{1 - (L_0/L_M)^{2/n}}{1 - (L_0)^{2/n}} \tag{2.5}$$

which is sometimes called the *Nagelkerke pseudo-R^2*.

Mixed-effects models There is no single analogue of R^2 for mixed models. The notion of "fraction of variance explained" becomes non-trivial, because mixed models contain several types of variance.[3] However, the likelihood ratio-based definition of R^2 ("net improvement from the baseline to the full model") can still be used, by simply plugging the intercept-only model and full model likelihoods into Eq. (2.4) for linear mixed models, and into Eq. (2.5) for mixed-effect logistic regression models (Kramer, 2005). These R^2 measures for mixed models serve as intuitive measures of model goodness relative to a baseline, and range between 0 and 1.

[3] For generalizations of R^2 to explain the amount of variance explained at different levels of mixed models, see Snijders and Bosker (2011: Ch. 7); Gelman and Hill (2007: Sec. 21.5).

CHAPTER 3

AUTOMATIC MEASUREMENT OF VOICE ONSET TIME

3.1 Introduction

Huge corpora of speech, both from laboratory or naturalistic settings, are becoming increasingly available and easy to construct, and promise to change the questions researchers can ask about human speech production.[1] However, this promise depends on the development of accurate algorithms to quicken or replace manual measurement, which becomes infeasible for large corpora. With a few important exceptions (such as pitch and vowel formants), such algorithms do not currently exist for most quantities that are widely measured in phonetic research. This chapter describes an automatic measurement algorithm for perhaps the most widely-measured consonantal variable, voice onset time.

Many stops, particularly in word-initial position, are produced with a burst. In phonologically voiced stops (/b/, /d/, /g/), the burst tends to be relatively short and voicing often begins before the burst. In phonologically voiceless stops (/p/, /t/, /k/), the burst tends to be relatively long, and voicing tends to begin after the burst. In pioneering work, Lisker and Abramson (1964, 1967) showed that a single parameter suffices to characterize much of the difference between voiced and voiceless stops, in perception and production: the time difference between the onset of the burst and the onset of voicing, which they termed voice onset time (VOT).[2] When voicing begins preceding the burst, VOT is negative and the stop is called *prevoiced*, or shows *voicing lead* (or *lead*). When voicing begins following the burst, VOT is positive and the stop shows *lag*.

VOT is measured in many clinical and non-clinical studies every year, requiring per-

[1] This chapter is a lightly-edited version of a submitted manuscript (Sonderegger and Keshet, 2012), which itself is a greatly expanded version of an earlier proceedings paper (Sonderegger and Keshet, 2010). Both are joint work with Joseph Keshet.

[2] As noted by Braun (1983), the concept of VOT goes back at least to the late 19th century (Adjarian, 1899). However, it was not widely known prior to its independent discovery by Lisker and Abramson.

haps thousands of transcriber-hours. Because controlling the duration of aspiration requires fine motor control, VOT has been extensively examined in speakers with communication disorders, both for diagnosis and treatment (Auzou et al., 2000). It is also widely measured in articulatory phonetics, for example to characterize how languages differ in the phonetic cues to stop contrasts (Cho and Ladefoged, 1999; Lisker and Abramson, 1964).

There have been a number of previous studies proposing algorithms for automatic VOT measurement. Previous work has used automatic measurements for speech recognition tasks (Ali, 1999; Niyogi and Ramesh, 1998, 2003; Stouten and van Hamme, 2009), phonetic measurement (Fowler et al., 2008; Tauberer, 2010), and accented speech detection (Hansen et al., 2010; Kazemzadeh et al., 2006). Some studies, like the current one, focus largely on the problem of VOT measurement itself, and evaluate the proposed algorithm by comparing automatic and manual measurements (Hansen et al., 2010; Lin and Wang, 2011; Stouten and van Hamme, 2009; Yao, 2009a). Our approach differs from all previous studies except one (Lin and Wang, 2011) in an important aspect. Instead of using a set of customized rules to estimate VOT, our system learns to estimate VOT from training data.

To replace manual measurement, we believe that an automatic VOT measurement algorithm should meet three criteria. Both because the burst and voicing onsets are often highly transient, and because the effects of interest (e.g., VOT difference between two conditions) in studies using VOT measurements are often very small, the algorithm should have high *accuracy* by the chosen measure of performance. The cues to the burst and voicing onset locations vary depending on many factors (speaking style, speaker's native language), and different labs have slightly different VOT measurement criteria. To account for such variation in the mapping between spectral/temporal cues and labeled VOT boundaries, the algorithm should be *trainable*: it should learn to measure VOT based on labeled data, and should perform well on diverse datasets. To meet the goal of replacing manual measurement, it should also be *adaptable* to a new dataset with little effort (i.e., training data).

This study proposes a supervised learning algorithm meeting all three criteria. The algorithm is trained on a set of manually-labeled examples, each consisting of a speech segment of an arbitrary length containing a stop consonant, and a label indicating the burst onset and the voicing onset, which we denote an *onset pair*. At test time the algorithm receives as input a speech segment containing a stop consonant, and outputs an onset pair and its corresponding VOT. The goal of the algorithm is to predict VOT as accurately as possible on unseen data.

Our algorithm belongs to the family of discriminative large-margin learning algorithms. A well-known member of this family is the support vector machine (SVM). The classical SVM algorithm assumes a simple binary classification task, where each input is of fixed length. The task of predicting VOT is more complex: the input is a speech segment of arbitrary length, and the goal is to predict the time between two acoustic events in the signal. Our algorithm is based on recent advances in kernel machines and large margin classifiers for structured prediction (Shalev-Shwartz et al., 2004; Taskar et al., 2003; Tsochantaridis et al., 2004). It maps the speech segment along with the target onset pair into a vector space endowed with an inner product. The vector space contains all possible onset pairs, and during training the algorithm tries to find a linear classifier that separates the target onset pair, as well as all "nearby" onset pairs (in terms of the cost function), from all other possible onset pairs in this vector space. At test time, the algorithm receives unseen speech segments. Each segment is mapped to the same vector space, and the most probable onset pair (and hence VOT) is predicted.

For this method to work and achieve high accuracy, the feature set must induce a vector space in which the target onset pair is both distinguishable and separable from other onset pairs. We achieve this by manually crafting a set of features which are informative about the precise locations of the burst and voicing onsets, and which tend to take on higher values for target onset pairs than for other onset pairs. The features leverage knowledge about how humans annotate VOT: using a variety of cues based on the spectrum, the waveform, and the output of speech processing algorithms (such as pitch trackers). We note that the feature sets typically used in speech recognition (e.g., MFCCs, PLPs) are not adequate for VOT measurement, since their time resolution is too coarse to accurately detect highly transient events such as burst onsets.

Another factor that controls the accuracy of the algorithm is the cost function used to evaluate how good a predicted VOT is, relative to its target value. Discriminative learning algorithms aim to maximize some measure of performance or minimize some cost function. The classic SVM, for example, is designed to minimize the zero-one loss function during training (i.e., the number of incorrect classifications). Our algorithm aims to minimize a special cost function, which is low if the predicted VOT is close to the manually-measured VOT and high otherwise. The function also does not penalize small differences between predicted and labeled VOT values during training, taking into account the fact that some measurement inconsistency (within or across annotators) is expected.

We evaluate our algorithm's accuracy in experiments on four datasets, using several methods to evaluate the algorithm's predictions relative to manual measurements. The datasets range across very different types of speech, testing the algorithm's applicability in different settings. To test the algorithm's adaptability to novel datasets where little or no labeled data is available, we perform experiments testing the algorithm's robustness to reducing the amount of training data, or training and testing on different datasets.

The remainder of the chapter is structured as follows. We first formally describe the problem of VOT measurement (Sec. 3.2), and describe our algorithm and the feature maps it takes as input (Sec. 3.3). We then turn to our experiments: first the datasets and evaluation methods used (Sec. 3.4), then experiments testing our method's accuracy (Sec. 3.5), and its robustness to decreasing the amount of training data and to mismatched train/test conditions (Sec. 3.6). We further evaluate our system by comparison with previous work

(Sec. 3.7), and by comparing regression models of variation in VOT induced by automatic and manual measurements. In Sec. 3.9 we sum up, and discuss directions for future work.

3.2 Problem setting

In the problem of VOT measurement, we are given a segment of speech, containing a stop consonant (plosive) followed by a voiced phone. The goal is to predict the time difference between the onset of the stop burst and the onset of voicing in the following phone. The speech segment can be of arbitrary length, but should include at most one burst, and its beginning need not be precisely synchronized with the stop's burst or closure; it is only required that the segment begins before the burst onset.

Throughout this chapter we write scalars using lower case Latin letters (x), and vectors using bold face letters ($\mathbf{x}$). A sequence of elements is denoted with a bar ($\bar{\mathbf{x}}$) and its length is written $|\bar{\mathbf{x}}|$.

We represent each speech segment by a sequence of acoustic feature vectors $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where each $\mathbf{x}_t$ ($1 \leq t \leq T$) is a D-dimensional vector. We denote the domain of the feature vectors by $\mathcal{X} \subset \mathbb{R}^D$. (The precise set of features used is described below.) Because different segments have different lengths, T is not fixed; we denote by $\mathcal{X}^*$ the set of all finite-length sequences over $\mathcal{X}$. Each segment is associated with an *onset pair* : $t_b \in \mathcal{T}$, the onset of the burst (in frames), and $t_v \in \mathcal{T}$, the onset of voicing of the following phone, where $\mathcal{T} = \{1, \ldots, T\}$. Given the speech segment $\bar{\mathbf{x}}$, our goal is to predict $t_v - t_b$: the length of time that passes between the beginning of the stop consonant's burst and the beginning of voicing in the following voiced phone. We assume here that $t_b < t_v$, and leave the case of "prevoiced" stops (where $t_b > t_v$), to future work. Our goal is to learn a function f from the domain of all speech segments $\mathcal{X}^*$ to the domain of all onset pairs $\mathcal{T}^2$.

3.3 Learning apparatus

In this section we describe a discriminative supervised learning approach for learning a function f from a training set of examples. Each example consists of a speech segment $\bar{\mathbf{x}}$ and a label (t_b, t_v). Our goal is to find a function that performs well on the training set, as well as on unseen examples. The performance of f is measured by the percentage of predicted VOT values, $t_v - t_b$, which are within a time threshold of the manually-labeled values.

Formally, given a speech segment $\bar{\mathbf{x}}$, let $(\hat{t}_b, \hat{t}_v) = f(\bar{\mathbf{x}})$ be the predicted onset pair. The *cost* associated with predicting $(\hat{t}_b, \hat{t}_v)$ when the manually-labeled pair is (t_b, t_v) is measured by a cost function, $\gamma : \mathcal{T}^2 \times \mathcal{T}^2 \to \mathbb{R}$. The function used in our experiments is of the form:

$$\gamma\left((t_b, t_v), (\hat{t}_b, \hat{t}_v)\right) = \max\{|(\hat{t}_v - \hat{t}_b) - (t_v - t_b)| - \epsilon, 0\}, \tag{3.1}$$

that is, only differences between the predicted VOT and the manually labeled VOT that are greater than a threshold ϵ, are penalized. This cost function takes into account that manual measurements are not exact, and ϵ can be adjusted according to the level of measurement uncertainty in a dataset. For brevity, we denote $\gamma = \gamma\left((t_b, t_v), (\hat{t}_b, \hat{t}_v)\right)$.

We assume that the training examples are drawn from $\mathcal{Q}$, a fixed (but unknown) distribution over the domain of the examples, $\mathcal{X}^* \times \mathcal{T}^2$. The goal of training is to find the f that minimizes the expected cost between predicted and manually-labeled VOT on examples from $\mathcal{Q}$, where the expectation is taken with respect to this distribution:

$$\mathbb{E}_{(\mathbf{x}, t_b, t_v) \sim \mathcal{Q}}\Big[\gamma\big((t_b, t_v), f(\bar{\mathbf{x}})\big)\Big].$$

Unfortunately, because we do not know $\mathcal{Q}$, we cannot simply compute this expectation. However, it still turns out to be possible to find f under lenient assumptions. We assume that our training examples are identically and independently distributed (i.i.d.) according to the distribution $\mathcal{Q}$, and that f is of a specific parameterized form. Below, we explain how to use the training set in order to find parameters of f which achieve a small cost on the training set, and a small cost on unseen examples with high probability as well.

We first describe the specific form used for the function f. Following the structured prediction scheme (Taskar et al., 2003; Tsochantaridis et al., 2004), f is constructed from a predefined set of N *feature maps*, $\{\phi_j\}_{j=1}^N$, each a function of the form $\phi_j : \mathcal{X}^* \times \mathcal{T}^2 \to \mathbb{R}$. That is, each feature map takes a speech segment $\bar{\mathbf{x}}$ and a proposed onset pair (t_b, t_v), and returns a scalar which, intuitively, should be higher if the onset pair makes sense given the speech segment, and should be lower if it does not. Each feature map can be thought of as an estimation of the probability of the onset pair given the speech segment (although the feature map need not actually be a proper probability distribution). For example, one feature map we use is the average energy of $\bar{\mathbf{x}}$ over frames in t_b to t_v, minus the average energy over frames in 1 to t_b. This feature map is expected to be high if t_b and t_v are located at the beginning and end of a stop burst following a closure, and low otherwise. Other feature maps might target the proper location of t_v or t_b (individually), or target VOT values ($t_v - t_b$) within a particular range. Note that the *features*, which the sequence $\bar{\mathbf{x}}$ is composed of, are oblivious to the locations of t_b and t_v, whereas the *feature maps* are specifically tailored to handle them.

Our VOT prediction function f is a linear function of the feature maps, where each feature map ϕ_j is scaled by a weight w_j. Linearity is not a very strong restriction, since the feature maps are arbitrary (so a non-linear dependency could be included as a further feature map). The overall score of an onset pair (t_b, t_v) is

$$\sum_{j=1}^{N} w_j \phi_j(\bar{\mathbf{x}}, t_b, t_v) = \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, t_b, t_v),$$

where we use vector notation for the feature maps, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$, and for the weights

$\mathbf{w} = (w_1, \ldots, w_N)$. Given $\bar{\mathbf{x}}$, f returns the onset pair which maximizes the overall score:

$$f(\bar{\mathbf{x}}) = \arg \max_{(t_b, t_v)} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, t_b, t_v), \tag{3.2}$$

In words, f gets as input a speech segment $\bar{\mathbf{x}}$ composed of a sequence of acoustic features, and returns a predicted onset pair by maximizing a weighted sum of the scores returned by each feature map ϕ_j.

We now describe the set of feature maps used (Sec. 3.3.1), then turn to how $\mathbf{w}$ is estimated from a training set of examples, so as to minimize the cost function defined in Eq. (3.1) (Sec. 3.3.2).

3.3.1 Features and feature maps

Consider the speech segment $\bar{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ consisting of T frames, where each acoustic feature vector $\mathbf{x}_t$ consists of D features. We extracted 7 ($D = 7$) acoustic features every 1 ms. The first 4 features refer to a short-time Fourier transform (STFT) taken with a 5 ms Hamming window: the log of the total spectral energy, E_{total}; the log of the energy between 50–1000 Hz, E_{low}; the log of the energy above 3000 Hz, E_{high}; and the *Wiener entropy*, H_{wiener}, a measure of spectral flatness:

$$H_{\text{wiener}}(t) = \log \int |P(f,t)|^2 df - \int \log |P(f,t)|^2 df,$$

where $P(f,t)$ is the STFT of the signal at frequency f and time t. The low frame rate and window size are used for fine time resolution, because the burst and voicing onsets are highly transient events.

The fifth feature, R_l, is extracted from the signal itself: the maximum of the FFT of its autocorrelation function, starting 6 ms before and ending 18 ms after the frame center. The sixth feature is the 0/1 output of a voicing detector based on the RAPT pitch

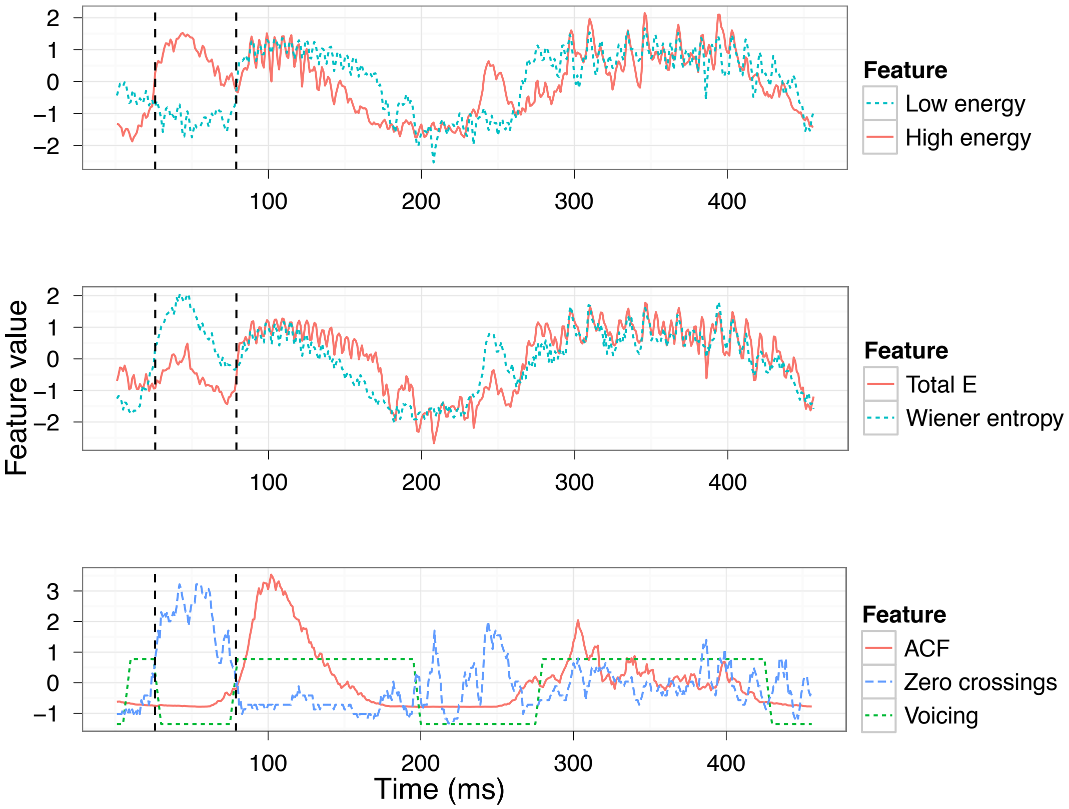

Figure 3.1: Values of the seven acoustic features for an example speech segment (the word "can't"). Vertical lines show the burst and voicing onsets.

tracker (Talkin, 1995), smoothed with a 5 ms Hamming window. The seventh feature is the number of zero crossings in a 5 ms window around the frame center. Fig. 3.1 shows the trajectories of the 7 features for one speech segment (the word "can't").

Before presenting the feature maps, we introduce notation for *local differences*. Let x^d be the d^{th} acoustic feature (of an arbitrary speech segment). $\Delta_t^s(x^d)$, the local difference of resolution s applied to the acoustic feature x^d, is defined as the difference between the mean of x^d over frames $\{t, \ldots, \min(t+s, T)\}$ and the mean of x^d over frames $\{\max(t-s, 0), \ldots, t\}$. This quantity provides a local approximation of the derivative of x^d at frame t, with resolution parametrized by s.

We now turn to the feature maps. For each example $(\bar{\mathbf{x}}, t_b, t_v)$, 61 feature maps ($N = 61$) were calculated. As described above, each feature map describes a scalar quantity that should be high for an onset pair which makes sense given the speech segment, and low otherwise. To ensure comparability of the values of feature maps across examples, each feature map was standardized (centered and divided by its standard deviation) within each example.

The feature maps are summarized in Table 3.1, where they are split into 7 types. We describe the intuition behind each type in turn.

> *Type 1:* We expect the correct t_b to occur at points of rapid increase in certain features, such as E_{high}, indicating the onset of turbulent airflow; at these points the corresponding local difference features (denoted by Δ in Table 3.1) spike. In the example (Fig. 3.1), E_{high} and H_{wiener} rapidly increase at the correct t_b. The inclusion of the values of some features (denoted by F in Table 3.1) at t_b helps rule out locations where a feature rapidly changes, but already has a high value.
>
> *Type 2:* Similar to Type 1, but for features expected to change rapidly at voicing onset. In the example, all features change rapidly near the correct t_v.
>
> *Type 3:* We expect R_l to not change during the burst (where there is no periodicity); hence the mean and maximum of its local difference over (t_b, t_v) should be low, as is the case in the example.
>
> *Type 4:* Similar to Type 3, but taking into account that periodicity can begin towards the end of the burst; hence the mean and maximum are calculated over $(t_b, t_v - 10)$.
>
> *Type 5:* Features indicating an aperiodic spectrum (E_{high}, H_{wiener}) should be much greater during the burst than before the burst. Hence, the difference between their mean in (t_b, t_v) and in $(1, t_b)$ should be large, as is the case in the example.
>
> *Types 6, 7:* Features indicating a noisy spectrum (E_{high}, H_{wiener}) should be uniformly low before the burst begins, and hence should have small mean and max values over $(1, t_b - 5)$. (An endpoint slightly before t_b is used because these features

Table 3.1: Summary of the 61 feature maps. The feature maps fall into several types described in the text, each of which is evaluated for some of the 7 acoustic features (one per column). ("V" and "ZC" in columns 7–8 stand for "voicing" and "zero crossings".) F in row i and column corresponding to feature x_j indicates that there is a feature map of type i for feature x_j; Δ indicates there are three feature maps of type i for the local difference of feature x_j, evaluated at $s = 5, 10, 15$. For example, the F, Δ in row 2, column 1 denotes four feature maps: $E_{\text{low}}(t_b)$, $\Delta^5_{t_b}(E_{\text{low}})$, $\Delta^{10}_{t_b}(E_{\text{low}})$, and $\Delta^{15}_{t_b}(E_{\text{low}})$.

Feature map type	E_{total}	E_{low}	E_{high}	H_{wiener}	R_l	V	ZC
1. Value at t_b	F, Δ	Δ	F, Δ	F, Δ	Δ		
2. Value at t_v	F, Δ	F, Δ	F, Δ	F, Δ	F, Δ	Δ	Δ
3. Mean & max over (t_b, t_v)					Δ		
4. Mean & max over $(t_b, t_v - 10)$					Δ		
5. Mean over (t_b, t_v) − mean over $(1, t_b)$		F		F			
6. Mean over $(1, t_b - 5)$		F		F		F	
7. Max over $(1, t_b - 5)$		F		F			

may already be rising by the burst onset.) We also expect there to be little voicing in this interval, and hence the voicing feature should have a low mean value.

The feature maps were chosen based on manual inspection of trajectories of the 7 acoustic features for labeled examples. They reflect knowledge about the effects of stop bursts and voicing on the spectrum, as well as knowledge about the process of VOT measurement itself. For example, feature types 3–4 take into account that the point labeled as the voicing onset can be somewhat later than the first signs of periodicity (Type 4), or synchronous with them (Type 3).[3]

3.3.2 *A discriminative algorithm*

We now describe a simple iterative algorithm for learning the weight vector $\mathbf{w}$, based on the *Passive-Aggressive* family of algorithms for structured prediction of Crammer et al. (2006), where the interested reader can find a more detailed description. Pseudocode for the algorithm is given in Fig. 3.2.

[3] Francis et al. (2003) discuss the relationship between voicing onset's true location and common criteria for annotating it based on the speech signal.

```
INPUT: training set $S = \{(\bar{\mathbf{x}}^1, t_b^1, t_v^1), \ldots, (\bar{\mathbf{x}}^m, t_b^m, t_v^m)\}$ ;
        parameters $C$ and $\epsilon$
INITIALIZATION: $\mathbf{w}^0 = \mathbf{0}$
FOR $\tau = 1, 2, \ldots$
  Pick example $(\bar{\mathbf{x}}^i, t_b^i, t_v^i)$ from $S$
  Predict
  $(\hat{t}_b^\tau, \hat{t}_v^\tau) = \arg\max_{(t_b, t_v)} \mathbf{w}^{\tau-1} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}^i, t_b, t_v) + \gamma((t_b, t_v), (t_b^i, t_v^i))$
  Set $\Delta\boldsymbol{\phi}^\tau = \boldsymbol{\phi}(\bar{\mathbf{x}}^i, t_b^i, t_v^i) - \boldsymbol{\phi}(\bar{\mathbf{x}}^i, \hat{t}_b^\tau, \hat{t}_v^\tau)$
  Set $\alpha^\tau = \min\left\{C, \frac{\gamma\left((t_b, t_v), (\hat{t}_b^\tau, \hat{t}_v^\tau)\right) - \mathbf{w}^{\tau-1} \cdot \Delta\boldsymbol{\phi}^\tau}{\|\Delta\boldsymbol{\phi}^\tau\|^2}\right\}$
  Set $\mathbf{w}^\tau = \mathbf{w}^{\tau-1} + \alpha^\tau \Delta\boldsymbol{\phi}^\tau$
OUTPUT: $\mathbf{w}^* = \sum_\tau \mathbf{w}^\tau$
```

Figure 3.2: Passive-Aggressive algorithm for training the VOT prediction function.

The algorithm receives as input a training set $S = \{(\bar{\mathbf{x}}^i, t_b^i, t_v^i)\}_{i=1}^m$ of m examples and a parameter C, and works in rounds. At each round, an example is presented to the algorithm, and the weight vector $\mathbf{w}$ is updated. We denote by $\mathbf{w}^\tau$ the value of the weight vector after the τ^{th} iteration. Initially we set $\mathbf{w}^0 = \mathbf{0}$. Let $(\hat{t}_b^\tau, \hat{t}_v^\tau)$ be the cost-adjusted prediction onset pair for the i^{th} example according to $\mathbf{w}^{\tau-1}$,

$$(\hat{t}_b^\tau, \hat{t}_v^\tau) = \arg\max_{(t_b, t_v)} \mathbf{w}^{\tau-1} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}^i, t_b, t_v) \, + \, \gamma\left((t_b, t_v), (t_b^i, t_v^i)\right) . \tag{3.3}$$

We set the weight vector $\mathbf{w}^\tau$ to be the minimizer of the following optimization problem,

$$\begin{aligned} &\min_{\mathbf{w}, \xi \geq 0} \; \frac{1}{2}|\mathbf{w} - \mathbf{w}^{\tau-1}|^2 + C\xi \\ &\text{s.t. } \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}^i, t_b^i, t_v^i) - \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}^i, \hat{t}_b^\tau, \hat{t}_v^\tau) \geq \gamma^\tau - \xi \; , \end{aligned} \tag{3.4}$$

where $\gamma^\tau = \gamma\left((t_b, t_v), (\hat{t}_b^\tau, \hat{t}_v^\tau)\right)$, C serves as a complexity-accuracy trade-off parameter (as in the SVM algorithm), and ξ is a non-negative slack variable which indicates the loss of the i^{th} example. Intuitively, we would like to minimize the loss of the current example (the slack variable ξ) while keeping the weight vector $\mathbf{w}$ as close as possible to our previous weight vector $\mathbf{w}^{\tau-1}$. The constraint ensures that the projection of the manually-labeled onset pair (t_b^i, t_v^i) onto $\mathbf{w}$ is higher than the projection of the predicted pair $(\hat{t}_b^\tau, \hat{t}_v^\tau)$ onto $\mathbf{w}$ by at least the cost function between them (γ^τ). It can be shown (Crammer et al., 2006) that the solution to the above optimization problem is

$$\mathbf{w}^\tau = \mathbf{w}^{\tau-1} + \alpha^\tau \Delta\boldsymbol{\phi}^\tau \; , \tag{3.5}$$

where $\Delta\boldsymbol{\phi}^\tau = \boldsymbol{\phi}(\bar{\mathbf{x}}^i, t_b^i, t_v^i) - \boldsymbol{\phi}(\bar{\mathbf{x}}^i, \hat{t}_b^\tau, \hat{t}_v^\tau)$. The value of the scalar α^τ, shown in Fig. 3.2, is based on the cost function γ^τ, the different scores that the manually-labeled onset pair and the predicted pair received according to $\mathbf{w}^{\tau-1}$, and a parameter C.

Given a training set of m examples we iterate over its elements, possibly M times (epochs), and update the weight vector $M \cdot m$ times. To classify unseen utterances, we use the average of $\{\mathbf{w}^1, \ldots, \mathbf{w}^{Mm}\}$, denoted by $\mathbf{w}^*$.

A theoretical analysis (Dekel et al., 2004; Keshet et al., 2007) shows that with high probability, the function learned using our algorithm will have good generalization properties: the expected value of the cost function when the algorithm is applied to unseen data is upper-bounded by the loss of the algorithm during training, plus a complexity term that goes to zero linearly with the number of training examples. For readers familiar with structural SVMs (Taskar et al., 2003; Tsochantaridis et al., 2004), we note that the same analysis suggests that the average loss of the Passive-Aggressive solution is comparable to the average loss of the structural SVM solution, while the structured Passive-Aggressive algorithm is much easier to implement and faster to train.

3.4 Experiments: Preliminaries

We first describe the datasets used in our experiments (Sec. 3.4.1), and the different methods used to evaluate the algorithm's performance (Sec. 3.4.2).

3.4.1 Datasets

Our experiments make use of four datasets, each consisting of audio of English words beginning with voiceless stops (/p/, /t/, /k/). For each word, the algorithm described above for training a VOT prediction function requires the VOT boundaries (t_b, t_v) and the word boundaries (where to begin and end searching for the VOT). We describe how VOT boundaries and word boundaries were annotated for each dataset, and briefly describe relevant aspects of each dataset.

The datasets vary along several dimensions, summarized in Table 3.2. Their speaking styles range in naturalness, from isolated words to read sentences to conversational

Table 3.2: Comparison of datasets used in experiments. A=American, B=British, L1=first-language, L2=second-language.

Dataset	Style	Dialect	Environment	N
TIMIT	read sentences	A	laboratory	5535
BB	conversational	B	TV studio	704
SWITCHBOARD	conversational	A	telephone	893
PGWORDS	isolated words	A (L1, L2)	laboratory	6795

speech. Three broad types of English accents are represented (American, British, Portuguese-accented). Finally, the recording conditions vary greatly. We perform experiments on several different datasets in order to test the robustness of our approach. The promise of learning a function to measure VOT is that it should perform well on diverse datasets because it can be re-trained on data from each one. By using several datasets, we show empirically that this is the case.

TIMIT The TIMIT corpus (Garofolo et al., 1993) consists of segmentally-transcribed sentences read by 630 American English speakers from 8 dialect regions. It is widely used by speech recognition researchers, and to a lesser extent by phoneticians (e.g., Keating et al., 1994). The TIMIT transcriptions are phonetic rather than phonemic, and there are two phone labels corresponding to each stop phoneme (/p/, /t/, /k/, /b/, /d/, /g/), for the closure and burst (e.g., pcl and p for /p/). Thus, each underlying stop can be annotated as a closure alone, a burst alone, a closure and burst, a different phone altogether, or nothing (if it is deleted). We restrict ourselves to all words (excluding the SA1 and SA2 utterances)[4] transcribed as beginning with an unvoiced stop burst (either preceded by a closure or not), followed by a voiced segment; this results in 5535 tokens, from all 630 speakers.

[4] These correspond to two sentences which are produced by all speakers in TIMIT; all other utterances correspond to sentences which are produced by a subset of speakers, or a single speaker. Thus, there are many more stops from SA1 and SA2 utterances than from other utterances, and we risk overfitting to stops from SA utterances if they are not excluded. Omitting the SA utterances is a common step in studies which use TIMIT, to prevent overfitting.

The VOT boundaries (t_b, t_v) were taken to be the burst boundaries from the TIMIT transcription. Because the burst sometimes ends after the onset of voicing, this step is an approximation, one which allows us to take advantage of the size of TIMIT, and test our algorithm on a widely used dataset. The word boundaries were also taken from the TIMIT transcription, except for some pathological cases where a word boundary coincided with the beginning or end of the burst. For words annotated as beginning with only a burst (and no closure), the left word boundary was taken to be 50 ms before the burst onset. For words annotated as consisting solely of an unvoiced stop (e.g., "to" transcribed as `tcl t`), the right word boundary was taken to be 25 ms after the end of the burst. These corrections were made because our algorithm assumes that the burst and voicing onsets lie within the host word.

Big Brother (BB) This dataset comes from spontaneous speech in the Big Brother corpus examined in this thesis. Sound quality is generally very good (see Sec. 5.2). The data are a subset of 704 tokens from the manually-annotated corpus of VOTs for word-initial voiceless stops described in Sec. 5.2.2, annotated by (one of) two transcribers. The end of each word has also been annotated. Because the beginnings of words have not been annotated, we took the left word boundary of each word to be 25 msec before the burst onset. Stops with no following voiced segment were kept if there was still abrupt spectral change at the end of the burst, and excluded otherwise.

Switchboard The Switchboard corpus (Godfrey and Holliman, 1997) consists of spontaneous speech from telephone conversations between American English speakers. We chose subsets of 8 conversations, corresponding to 16 speakers. VOTs for all word-initial voiceless stops in these subsets were manually annotated by one transcriber if a burst was present (e.g., the stop was not realized as a flap), resulting in 893 examples. The boundaries of each word were manually determined. When a word boundary coincided with the burst or voicing onset (e.g., for a word realized as an isolated stop, with no following voicing), the word boundary was adjusted slightly left or right (for the left or right word boundaries, respectively), because our algorithm assumes that the burst and voicing onsets lie within the host word.

Paterson/Goldrick words (PGWORDS) This corpus consists of data from a laboratory study by Nattalia Paterson and Matt Goldrick (Paterson, 2011), investigating VOT in the speech of American English monolinguals and Brazilian Portuguese (L1)-English bilinguals. In this study, each of 48 speakers (24 monolinguals, 24 bilinguals) produced 144 isolated words, each beginning with a stop (/p/, /t/, /k/, /b/, /d/, /g/), in a picture naming task. Productions other than the intended label as well as those with code-switching or disfluencies were excluded. The VOT of each remaining word was manually measured by a single transcriber. We consider a subset of 6795 VOTs from this data, only from words beginning with voiceless stops. Because this dataset consists of words spoken in isolation, the choice of word boundaries is somewhat arbitrary. We took the left boundary to be 50 ms before the burst onset and the right boundary to be when the next prompt was given to the subject (1–3 seconds later).

3.4.2 Evaluation methods

There is no single obvious method for evaluating the performance of an automatic VOT measurement algorithm. Several methods have been used in previous work on automatic measurement, all based on the degree of discrepancy between automatic and manual measurements. Below, we measure our algorithm's performance by three methods: pure automatic/manual measurement discrepancy, comparison of automatic/manual discrepancy to intertranscriber agreement, and comparison of statistical models fit to automatic and manual measurements. We now describe each method and its motivation.

Distribution of automatic/manual difference The most common evaluation method used in previous work is examination of the distribution of differences between automatic and manual VOT measurements across a dataset. The algorithm's performance can then be reported either as the full (empirical) CDF of automatic/manual differences (as in Stouten and van Hamme, 2009), or as the percentage of examples with automatic/manual difference below some fixed set of values, the *tolerances* (as in Lin and Wang, 2011) . Reporting statistics about the CDF of automatic/manual differences is a standard evaluation method in ASR tasks, such as forced alignment of phoneme sequences, where the goal is to predict the location of boundaries in a speech segment (e.g., Brugnara et al., 1993; Keshet et al., 2007). The full CDF gives a more complete picture of experimental performance, while performances at fixed tolerances are easier to interpret. We usually report results in both ways below when discussing the distribution of automatic/manual differences for an experiment (e.g., Fig. 3.3, Table 3.3).

Two other evaluation methods, where a single measure of error is calculated from the set of automatic/manual differences, have also been used in previous work. These are discussed in Sec. 3.7, where they are used to compare our algorithm with previous work.

Comparison to interrater reliability A disadvantage of evaluation using the distribution of automatic/manual differences is that it is not clear what the gold standard is. VOT measurements for the same example are expected to vary somewhat between transcribers, or within a single transcriber (measuring at different times). Intuitively, progressively better automatic/manual agreement is good up to a point, but automatic/manual agreement which is *too* good means overfitting to the particular set of manual measurements. One solution is to compare the automatic/manual CDF to interrater reliability (IRR): a CDF of differences between two transcribers' VOT measurements. In this view, the gold standard is for automatic and manual measurements to agree as well as two sets of manual measurements from the same dataset. We compare the predicted/manual difference CDF to an IRR CDF for experiments on all datasets where IRR data is available (BB, SWITCHBOARD, PGWORDS).

Model-based comparison Our last evaluation method is more directly related to the setting in which VOT is usually measured: phonetic studies addressing clinical or theoretical questions. In such studies, the question is how some predictor variables—such as the stop consonant's place of articulation, or whether the speaker has Parkinson's disease—affects VOT across a dataset. This is typically assessed by performing a statistical analysis (such as ANOVA or multiple linear regression) of the effect of the predictors on VOT, and reporting the statistical significance and values of model parameters of interest. Thus, to test whether automatic VOT measurements from our algorithm can be used to replace manual measurements, a sensible test is to compare the *statistical models* induced from automatic and manual measurements for the same dataset, rather than directly comparing the automatic and measurements of individual tokens (as in the evaluation methods described above). The goal is for the values and statistical significances of the two models to be as similar as possible. We note that good performance on this model-based evaluation method does not trivially follow from good performance on an evaluation method based on individual tokens, or vice versa.

3.4.3 *Experiments: Overview*

The next four sections describe a series of experiments to evaluate the algorithm's performance, using each of the evaluation methods just described. In Sec. 3.5 we describe experiments using the full amount of data available, and where training and testing data are (disjoint subsets) from the same dataset; we call these *base experiments*. We then (Sec. 3.6) evaluate the robustness of the results obtained in the base experiments to decreasing the amount of training data, or training and testing on different datasets. In Sec. 3.7 we compare our algorithm to previous work as closely as possible. Finally, we evaluate the algorithm by model-based comparison (Sec. 3.8).

In all experiments, we only considered burst onsets t_b within 0–150 ms of the start of the word, and voicing onsets t_v 15–200 ms later than t_b; this step attempts to restrict the algorithm's focus to the first two segments of each word.

3.5 Experiments I: Base

The evaluation method used for the base experiments is simply the distribution of automatic/manual differences. Where IRR data is available (all datasets except TIMIT), this distribution is compared to the distribution of differences between transcribers.

The structure of each base experiment was the same: the dataset was split into training, development, and test sets corresponding to subsets of speakers. The parameters C, ϵ, and M (number of epochs) were tuned by training a weight vector on the training set for each parameter triplet in the ranges $C \in 0.01, 0.1, 1, 5, 10, 100$, $\epsilon \in \{2, 3, 4, 5\}$, and $M \in \{1, 2, 3, 4, 5\}$ (for TIMIT, PGWORDS) or $M \in \{1, 2, 3, 4, 5, 6, 7, 9, 11, 15\}$ (for BB, SWITCHBOARD).[5] The weight vector was selected that gave the lowest mean absolute difference between predicted and actual VOT over examples in the development set. This $\mathbf{w}^*$ was then applied to predict VOTs for examples in the test set.

For each experiment, Fig. 3.3 and Table 3.3 summarize the distribution of automatic/manual differences over the test set, and the distribution of intertranscriber differences over the set of double-transcribed examples (except for TIMIT).

3.5.1 Big Brother

For the BB dataset, the training/development/test sets consisted of 405/142/160 examples (2/1/1 speakers), and the parameter values chosen by tuning on the development

[5] For the two larger datasets (TIMIT and PGWORDS), experiments for $M > 5$ took prohibitively long, and it was clear that performance worsened for $M > 2$.

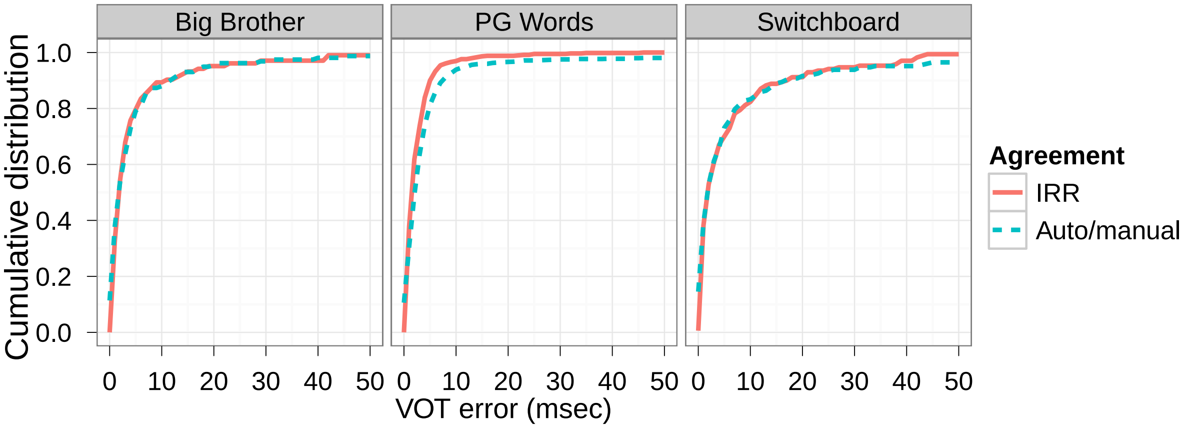

Figure 3.3: Performance on BB, SWITCHBOARD, and PGWORDS datasets in base experiments (Sec. 3.5). For each dataset, distributions are given of absolute differences between automatic and manual measurements, and between two human transcribers (IRR).

set were $C = 5$, $\epsilon = 3$, and $M = 15$. A subset of the data (108 stops; 15.3%) was double-transcribed, by two independent transcribers. (Neither one trained the other, and there was no attempt made to synchronize transcription criteria.)

In comparison to IRR, the algorithm performs very well: the automatic/manual curve and the IRR curve are essentially the same. That is, the automatic measurements match manual measurements as well as two human transcribers match each other.

3.5.2 *Switchboard*

For the SWITCHBOARD dataset, the training/development/test sets consisted of 563/102/288 examples (6/1/2 speakers), and the parameter values chosen by tuning on the development set were $C = 1$, $\epsilon = 5$, and $M = 5$. A subset of the data (171 stops; 19.1%) was double-transcribed, by two semi-independent transcribers. (One transcriber had trained the other about one year previously, on a different dataset, but no attempt at synchronizing transcription criteria was made for the SWITCHBOARD data.)

The algorithm again performs very well, by comparison to IRR: automatic/manual differences are slightly lower than inter-transcriber differences at tolerances up to about

Table 3.3: Performance in base experiments (Sec. 3.5), given as percentage of examples in the test set with automatic/manual difference (for all datasets) or inter-transcriber difference (for all datasets except TIMIT) below a series of fixed tolerance values. (For example, 46.1% of examples in the TIMIT test set had automatic and manual measurements differing by $\leq$2 ms.)

Dataset	Experiment	≤ 2 ms	≤ 5 ms	≤ 10 ms	≤ 15 ms	≤ 25 ms	≤ 50 ms
BB	Auto/manual	53.5	79.3	88.1	93.1	96.2	98.7
	Intertranscriber	54.4	79.6	89.3	93.2	96.1	99.0
SWITCHBOARD	Auto/manual	53.1	73.3	83.3	89.0	93.4	96.5
	Intertranscriber	52.9	70.0	82.4	88.8	94.1	99.4
PGWORDS	Auto/manual	49.1	81.3	93.9	96.0	97.2	98.1
	Intertranscriber	61.9	90.0	96.9	98.6	99.5	100.0
TIMIT	Auto/manual	46.1	67.2	85.0	94.7	98.1	99.0

20 ms, and slightly higher above 20 ms, becoming significantly higher above 40 ms.

3.5.3 *Paterson/Goldrick words*

For the PGWORDS dataset, the training/development/test sets consisted of 4151/403/1340 examples (28/3/6 speakers), and the parameter values chosen by tuning on the development set were $C = 10$, $\epsilon = 5$, and $M = 2$. A subset of the data (591 stops; 7.3%) was double-transcribed, by two transcribers. (One transcriber trained the other, and they worked together on synchronizing measurement criteria for this dataset.)

The algorithm performs less well on this dataset than for BB or SWITCHBOARD, by comparison to IRR: automatic/manual VOT measurement differences are higher than inter-transcriber differences, at all tolerances. A possible explanation for this difference in performance is that the intertranscriber data for the three datasets are not comparable. The transcribers for PGWORDS worked together to synchronize their measurement criteria on this dataset, while the transcribers for BB and SWITCHBOARD did not. Thus, the algorithm's performance on PGWORDS might be closer to IRR if intertranscriber data were used from two independent transcribers.

3.5.4 TIMIT

For the TIMIT dataset, we used Halberstadt's (1998) split of speakers into training, development, and test sets (specifically, "full" test) consisting of 4132/397/1006 examples (462/50/118 speakers). The parameter values chosen by tuning on the development set were $C = 5$, $\epsilon = 4$, and $M = 2$.

Performance for TIMIT is worse than other datasets (greater automatic/manual differences) for tolerances up to about 10 ms, and slightly better at tolerances above 10 ms. However, it is not clear how comparable the results for different datasets are, given that the TIMIT annotations actually denote burst boundaries rather than VOT. Below (Sec. 3.7) we will more directly evaluate our TIMIT results, by comparing them with previous work on automatic VOT estimation that also uses test data from TIMIT.

3.6 Experiments II: Robustness

The base experiments show that our algorithm generally performs very well on several datasets, evaluated against IRR; we show below that it also performs well relative to previous work (Sec. 3.7). However, in both cases we assume ideal training conditions: a relatively large training set is available to train $\mathbf{w}^*$, and the training and test sets consist of examples from the same corpus. In contrast, the typical use case for a VOT measurement algorithm is a corpus where little or no annotated data is available. For our algorithm to be practically useful, we must test how performance varies as these conditions are relaxed. If relatively few examples are needed to train $\mathbf{w}^*$, other researchers can annotate a small subset of data to train our algorithm; if performance varies little when the training and test corpora are not the same, researchers working on a new dataset can use one of our weight vectors pre-trained on a large corpus. This section presents experiments testing the algorithm's robustness to decreasing the amount of training data (Sec. 3.6.1), and to mismatched training and testing datasets, where a weight vector trained on one corpus is used to measure VOT for data from a different test corpus (Sec. 3.6.2).

3.6.1 Varying the amount of training data

For each of the base experiments, we tested the robustness of our algorithm to decreasing the amount of training data, holding the test set constant, as follows. Let N be the size of the training set for a given experiment. We chose a series of percentages p of the test set, such that pN spanned the range $(0, N]$, including N ($p = 1$). (We did not use the same values of p for each dataset because N varies greatly across our datasets.) For each p we chose a random subset of the training set of size pN and re-ran the experiment, using the same test set as the original ($p = 1$) experiment, and using the same parameter values (for C, ϵ, and M) as in the original experiment. Since the results depend on the particular subset of pN chosen, this procedure was repeated 25 times for each p.

The results of the experiments are summarized in Fig. 3.4. To focus on how performance changes as the amount of training data is decreased, we show results only for a subset of tolerances (and do not show the full CDFs of automatic/manual differences). Each point and its associated errorbars represent the mean and ± two standard deviations of the 25 runs at a fixed amount of training data.

For all datasets, the algorithm's performance is extremely robust to decreasing the amount of data. Performance stays essentially constant (mean performance within 2σ of performance at the highest N) (error bars overlap with those for the full training set) —until the amount of data is decreased below 250 training examples for PGWORDS and TIMIT, and below 25–50 examples for SWITCHBOARD and BB. Performance decreases more at lower tolerances (2–5 ms) than at higher tolerances (10–50 ms), especially for TIMIT and PGWORDS. To speak more quantitatively, we can focus on performance at 10 ms tolerance, shown in Table 3.4. Across all datasets, training on just 25 examples decreases performance at this tolerance by 1.8%–5.2%. These results suggest our algorithm can be quickly adapted to a new dataset with little training data.

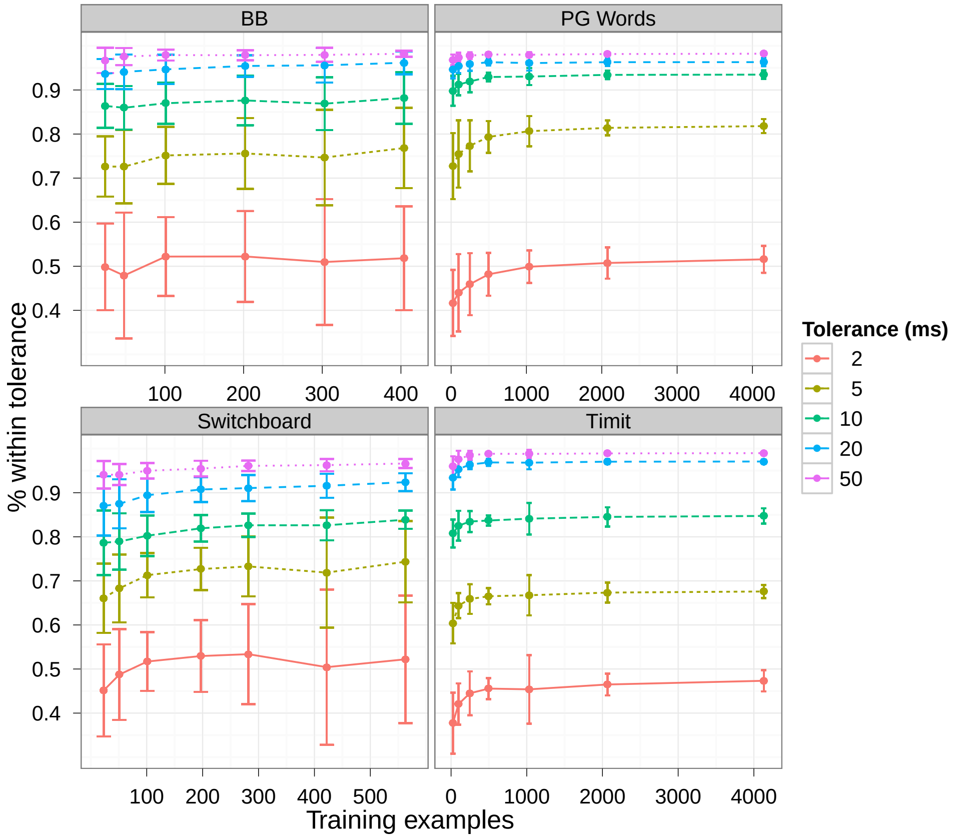

Figure 3.4: Results for experiments varying the amount of training data (Sec. 3.6.1): Percentage of tokens with automatic/manual difference below tolerance values (2, 5, 10, 20, 50 msec) as the amount of training data is varied. Points and errorbars indicate means ± 2 standard deviations across 25 runs.

Table 3.4: Mean performance at 10 ms tolerance (across 25 runs) in experiments where the amount of training data is decreased (Sec. 3.6.1), with number of training examples in parentheses, for the lowest number of training examples (n_1), the highest number of training examples (N), and the lowest number of training examples (n_2) for which performance is within 2σ of mean performance with the highest number of training examples.

Dataset	n_1	n_2	N
BB	86.4 (24)	86.4 (24)	88.2 (404)
SWITCHBOARD	78.6 (23)	78.6 (23)	83.8 (563)
TIMIT	80.8 (25)	82.5 (99)	84.7 (4132)
PGWORDS	89.8 (25)	91.2 (100)	93.5 (4151)

Table 3.5: Mean performance at 10 ms tolerance in experiments with mismatched training and test corpora (Sec. 3.6.2).

	Test corpus			
Training corpus	BB	SWITCHBOARD	TIMIT	PGWORDS
BB	88.0	79.4	75.6	77.2
SWITCHBOARD	84.3	83.3	78.0	86.9
TIMIT	81.1	81.1	84.8	84.9
PGWORDS	77.4	74.1	83.9	93.9

3.6.2 *Mismatched training and test corpora*

We now test how the algorithm's performance varies when different training and testing corpora are used. To compare to the base experiments (where the training set and test set were drawn from the same corpus), we conduct 12 additional experiments, corresponding to all possible choices of two distinct datasets for training and testing $\mathbf{w}^*$. We denote the weight vectors trained on each corpus in the base experiments as $\mathbf{w}^*_{\text{TIMIT}}$, etc. The weight vector for each corpus is applied to give automatic measurements for examples in the test sets of the other three corpora.

The distributions of automatic/manual differences for each pair of training and testing corpora are shown in Fig. 3.5. To discuss performance differences quantitatively, it will again be helpful to refer to performances at 10 ms for each curve, given in Table 3.5.

We note some patterns in these results. First, examining the full CDFs, it is always the case that the best performance for a test set from a given corpus is achieved using training data from that corpus. (This is visually clear except for the TIMIT test set, where the CDF corresponding to $\mathbf{w}^*_{\text{TIMIT}}$ is in fact higher than the CDF corresponding to $\mathbf{w}^*_{\text{PGWORDS}}$ at all tolerances.) Better performance when training and test data are drawn from the same distribution is not surprising, but it is useful to investigate how much performance drop to expect, for potential applications of the algorithm where re-training on data drawn from the same distribution as the test corpus would not be possible. (For example, real-time VOT detection in a novel recording environment)

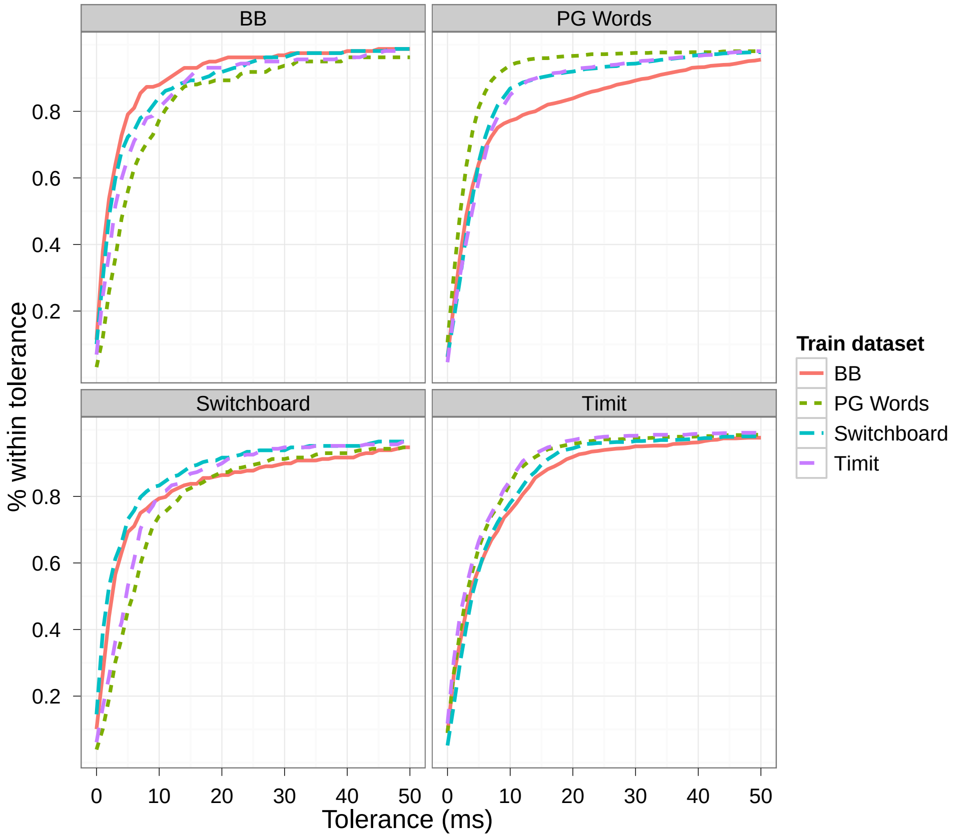

Figure 3.5: Distribution of automatic/manual differences as a function of the dataset the training set is drawn from, holding the test set constant. Lines for mismatched train/test datasets correspond to experiments described in Sec. 3.6.2. Lines for same train/test datasets correspond to the base experiments (Sec. 3.5).

How performance changes when different training and test corpora are used depends largely on the test corpus. For the BB, TIMIT, and SWITCHBOARD test sets, there is significant variance in how much using a different test corpus affects performance (1–11% at 10 ms), with some mismatched train/test pairs achieving performance near the corresponding matched train/test conditions in certain tolerance ranges. Performance on the PGWORDS test set is more dramatically affected by training on a different corpus, with a 7-17% performance drop at 10 ms depending on the training set. It is not clear why changing the weight vector used matters more for PGWORDS, compared to the other datasets.

3.6.3 Discussion

In this section, we have described experiments testing the robustness of our algorithm's performance to decreasing the amount of training data, and on mismatched training and test conditions. We found that performance is very robust to decreasing the amount of training data, and that the effect of mismatched training and test datasets depends on the particular datasets used. The motivation for these experiments was to determine whether our algorithm can be practically useful in applications to *new* datasets, without manually labeling a large amount of data. Our results suggest a positive answer: only a small number of manually-labeled VOTs (<250) are needed for training, in addition to a small number for validation (perhaps 50–200), to achieve near-maximum performance.

3.7 Evaluation I: Comparison with previous work

In this section, we compare our algorithm's performance with all previous studies on automatic VOT measurement (to our knowledge) that have examined agreement between automatic and manual measurements. We are able to compare directly (testing on the same test set) to two previous approaches (Lin and Wang, 2011; Stouten and van Hamme, 2009), and indirectly to two other approaches (Hansen et al., 2010; Yao, 2009a).

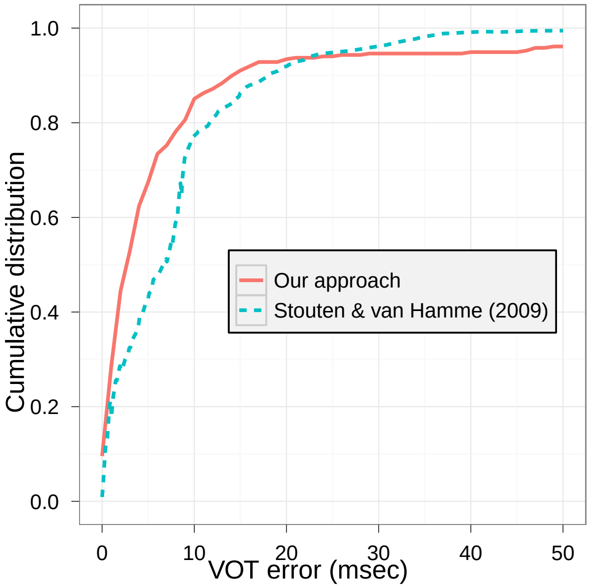

Figure 3.6: Comparison of automatic/manual differences for test data from Stouten and van Hamme (2009), using their algorithm and using our approach.

3.7.1 *Stouten and van Hamme (2009)*

Stouten and van Hamme (SvH) automatically measure VOT for stops from TIMIT, using a knowledge-based algorithm operating on time-frequency reassigned spectrograms. They consider voiced and voiceless stops, in all positions. They performed manual measurements for a subset of 582 stops (the “manual” dataset), and compared these to their automatic measurements.

We applied our algorithm to the 293 voiceless stops from SvH’s “manual” dataset, using $\mathbf{w}^*$ from the TIMIT base experiment. Because we are now not dealing only with stops in initial position, the left boundary where the algorithm begins searching for t_b was determined differently from our earlier TIMIT experiments. Each example was taken to start at the beginning of the segment preceding the burst (i.e., the closure, if one was present), and end at the right word boundary, where the segment and word boundaries were taken from TIMIT.

Fig. 3.6 and Table 3.6 show the distribution of automatic/manual differences, rela-

Table 3.6: Comparison of results for test data from Stouten and van Hamme (2009), using their algorithm and using our approach.

Algorithm	≤ 2 ms	≤ 5 ms	≤ 10 ms	≤ 15 ms	≤ 25 ms	≤ 50 ms
Stouten and van Hamme (2009)	28.6	42.8	77.0	86.2	**94.7**	**99.5**
Our approach	**44.6**	**67.6**	**85.1**	**91.1**	94.1	96.1

tive to SvH's manual measurements, for the two automatic measurement methods: our method (with $\mathbf{w}^*$ from the TIMIT base experiment) and SvH's method. (The error distribution is taken from their Fig. 5, using voiceless stops only.) Our method performs better for tolerances below about 22 ms, corresponding to 94% of examples. We note that our algorithm is at a significant disadvantage in this comparison: it was only trained on initial stops, but tested in stops in all positions, and the training and testing data were labeled by different annotators.

3.7.2 Lin and Wang (2011)

Lin and Wang automatically measure VOT for word-initial stops from TIMIT, using a multi-step process. Their approach makes use of two tools: (1) an HMM-based forced aligner, at either the state or phone level (they try both), using MFCC features; (2) two random forest detectors, trained to detect the onset of a burst phone and the onset of a voiced phone. For a given utterance, the forced aligner is first applied to the TIMIT phone transcription to find the approximate location of the burst for each stop. For a given stop, the random forest detectors are then deployed to find the burst and voicing onsets, possibly selecting from several candidates. If no candidates are found for an onset, the force-aligned burst boundary is used instead. The burst boundaries from the TIMIT annotation are used as a proxy for VOT (as we also did in our TIMIT experiments above).

Lin and Wang test their algorithm on 2344 word-initial stops from the TIMIT `test` set, of which 1174 are voiceless. We applied our algorithm to the voiceless stops, using the $\mathbf{w}^*$ from our TIMIT experiments above. Because this set of voiceless stops forms a subset of

Table 3.7: Comparison of results for test data from Lin and Wang (2011), using their algorithm and using our approach.

	< 5 ms	< 10 ms	< 15 ms	< 20 ms
Lin & Wang (2009)	58.9	80.7	90.5	94.2
Our approach	**60.5**	**81.4**	**93.0**	**96.8**

the complete TIMIT test set used in our experiments above, the word boundaries for each example have already been determined.

Table 3.7 shows performance on the test set for Lin and Wang's method (their Table IV, "voiceless" row) and our method. (Results are shown in the format used in their paper.) Our algorithm gives better performance at all tolerances, on average by 1.8%.

There are two differences between our setup and that of Lin and Wang which bear mentioning. First, as Lin and Wang note with respect to results reported in an earlier version of the current study (Sonderegger and Keshet, 2010), our training set contains data from many more TIMIT speakers than theirs (462 speakers in our setting versus 4 speakers in theirs). While the exact numbers of speakers used in the two approaches are not comparable because of differences in the training procedures,[6] we have shown above (Sec. 3.6.1) that our method is very robust to decreasing the amount of training data.

Second, Lin and Wang use forced alignment of the TIMIT phone transcription to determine the approximate region to search for the burst and voicing onsets for a given stop, while we simply assume that the word boundaries are known (using the TIMIT transcription); it is possible that our results would worsen using force-aligned word boundaries instead. However, experiments on the TIMIT test set reported in our earlier paper (Sonderegger and Keshet, 2010) found no performance difference, suggesting that assuming word boundaries are known is not responsible for our results.

[6] Our algorithm only considers portions of the speech signal in the training utterances from words beginning with initial voiceless stops; Lin and Wang's random forest detectors are trained using *all* of the training utterances. Thus, we believe the number of speakers whose utterances are used in training is not a good method for comparing the amount of training data used in the two approaches.

Table 3.8: Base experiments performance evaluated by metrics used by Yao (2009a) and Hansen et al. (2010).

	RMS t_b error (ms)	< 10% VOT error
TIMIT	6.5	66.5
BB	5.8	73.4
SWITCHBOARD	10.5	68.7
PGWORDS	6.8	81.2
Yao (2009a)	10.8	—
Hansen et al. (2010)	—	74.9

3.7.3 Yao (2009a); Hansen et al. (2010)

Our results can be compared less directly with two other studies where automatic and manual measurements are compared. Yao (2009a) determines VOT for initial voiceless stops in the Buckeye Corpus, using MFCC "spectral templates" in a knowledge-based algorithm. The evaluation metric used is the RMS error for automatic measurement of t_b only. Hansen et al. (2010) determines VOT for initial voiceless stops from CU-ACCENT, a corpus of laboratory speech consisting of single-word productions by native and non-native English speakers. They measure VOT with a knowledge-based algorithm, acting on the Teager Energy Operator representation of the speech signal. The evaluation metric used is the percentage of stops where the automatic measurement differs by <10% from the manual measurement.

Table 3.8 gives these metrics for the experiments on our four datasets reported above. With the caveat that comparison is difficult because of the different datasets used, our experiments' performance on these metrics compares favorably to previous work. All experiments have RMS t_b error less than Yao (2009a). Our best-performing experiment, PGWORDS, does better than Hansen et al. (2010) on the <10% VOT error metric. Importantly, the PGWORDS dataset is arguably the most comparable to Hansen et al.'s CU-ACCENT dataset: both consist of single-word productions in laboratory conditions, produced by both native and non-native speakers.

3.8 Evaluation II: Regression model comparison

We now evaluate the algorithm by comparing the regression models induced by automatically and manually-measured data, as described above (Sec. 3.4.2), for the PGWORDS dataset. We model two well-documented patterns of variation in VOT in this dataset. VOT is affected by the stop consonant's place of articulation, with the expected pattern /p/</t/</k/; (e.g., Cho and Ladefoged, 1999). VOT is also affected by the speakers linguistic knowledge. Bilinguals who speak languages with contrasting VOT systems produce distinct VOT patterns from monolingual speakers, reflecting the interaction of their two linguistic systems. When speaking English, bilinguals whose native language contrasts prevoiced vs. short lag VOTs produce shorter VOTs than those of English monolinguals (e.g., Fowler et al., 2008). Thus, we expect the bilingual Portuguese-English speakers to have lower VOT than the monolingual English speakers in the PGWORDS data.

3.8.1 Models

For each set of measurements (automatic, manual), we build a mixed-effects linear regression model of how VOT depends on two input variables: the speakers's language background, and the place of articulation of the initial consonant of the host word.[7] The models contain fixed effects for 3 predictors: language background (2 levels, dummy-coded; base level=Portuguese/English bilingual), stop place of articulation (3 levels, dummy-coded; base level=/p/), and the interaction of the two. By-speaker and by-word random intercepts are also included. Thus, VOT is modeled as varying by place of articulation and by language background, and the effect of place of articulation may differ between L1 English and L2 English speakers. Individual speakers and words have characteristic VOT values, normally distributed around the mean.

We constructed the automatic model and the manual model as follows. The PGWORDS

[7] Background on mixed-effects regression is given in Chapter 2; the particular type of model used here is described in Sec. 2.2.1.2.

dataset was split into a split and heldout set, such that split contained a random 75% of data points for each speaker, and heldout contained the remaining 25%. An automatic measurement was assigned to each data point in split by applying our algorithm (with the values of C, M, and ϵ used for the PGWORDS base experiment), using four-fold cross validation. (Speakers were split randomly into four groups; for each group, automatic VOT measurements were computed using $\mathbf{w}^*$ trained on data from the other 3 groups.) Each data point in split now had one automatic and one manual measurement, resulting in two datasets, differing only in whether VOT (y_i, in the notation above) was measured automatically or manually.

Each of the two datasets was trimmed for outliers, by discarding measurements further than 3 standard deviations within a speaker. For each dataset, a model of the form described above was fitted using the (Bates et al., 2011) package in R (Bates et al., 2011).

3.8.2 *Results*

We compare the automatic and manual models by comparing their fitted model parameters, and their predictions on held-out data.

Comparison of model parameters Table 3.9 shows the fixed-effect coefficient estimates ($\hat{\beta}$), their standard errors (SE($\hat{\beta}$), and corresponding p-value (for a Wald test applied to $\hat{\beta}$/SE($\hat{\beta}$)). The estimates and standard errors are extremely similar in the two models, with no fitted coefficient having a value in one model more than 1.25 standard errors away from its value in the other model. The significances of the fixed effect coefficients in the two models are also very similar: all coefficients are highly significant except the interaction of L1 with the first place of articulation contrast, which is marginal in the automatic model and not significant in the manual model.

The group means of the empirical data are /p/=67.9 ms, /t/=80.6 ms, /k/=79.7 ms for monolinguals; and /p/=41.3 ms, /t/=52.8 ms, /k/=65.2 ms for bilinguals. The fixed

Table 3.9: Summary of fixed effects in automatic and manual models: fixed effects coefficient estimates ($\hat{\beta}$), their standard errors, associated t statistic ($\hat{\beta}$/SE($\hat{\beta}$)), and p-values. *** denotes $p < 0.0001$.

	Automatic model				Manual model			
	$\hat{\boldsymbol{\beta}}$	SE($\hat{\boldsymbol{\beta}}$)	t	p	$\hat{\boldsymbol{\beta}}$	SE($\hat{\boldsymbol{\beta}}$)	t	p
Intercept	52.8	2.8	18.6	***	51.5	2.9	17.7	***
L1	11.0	2.4	4.6	***	10.6	2.5	4.3	***
POA=/t/	8.9	2.1	4.3	***	9.6	2.1	4.7	***
POA=/k/	17.1	1.7	10.3	***	17.7	1.7	10.6	***
L1:POA=/t/	1.4	0.8	1.8	0.072	0.79	0.72	1.1	0.27
L1:POA=/k/	-4.7	0.7	-7.0	***	-5.5	0.6	-8.8	***

effects for both models suggest that the trends observed in the empirical data are significant. English speakers have longer VOTs than bilingual Portuguese-English speakers and VOT depends on place of articulation as /p/</t/</k/, both expected results. There is also a significant interaction between L1 and phone. For bilingual speakers the entire /p/</t/</k/ pattern is significant. For monolinguals, the VOT difference between /p/ and /t/ is greater, and the VOT difference between /p/ and /k/ is smaller. The pattern of VOT dependence on place of articulation is thus closer to /p/</t/=/k/ for monolinguals, consistent with some previous studies of VOT in English word-initial voiceless stops (e.g., Cooper, 1991; Docherty, 1992).

Fig. 3.7 (left) plots the best linear unbiased predictors (BLUPs) of the random intercepts for each speaker in the two models, with horizontal and vertical errorbars corresponding to 95% confidence intervals in the manual and automatic models, respectively (1.96×standard errors). The models predict similar deviations from the mean for each speaker, with the confidence intervals for the automatic and manual by-word random intercepts overlapping for all but 1 of the 34 speakers. Fig. 3.7 (right) is an analogous plot for the random intercepts for each word. The models predict similar deviations from the mean for each word, with the confidence intervals for the automatic and manual by-word random intercepts overlapping for all 206 words.

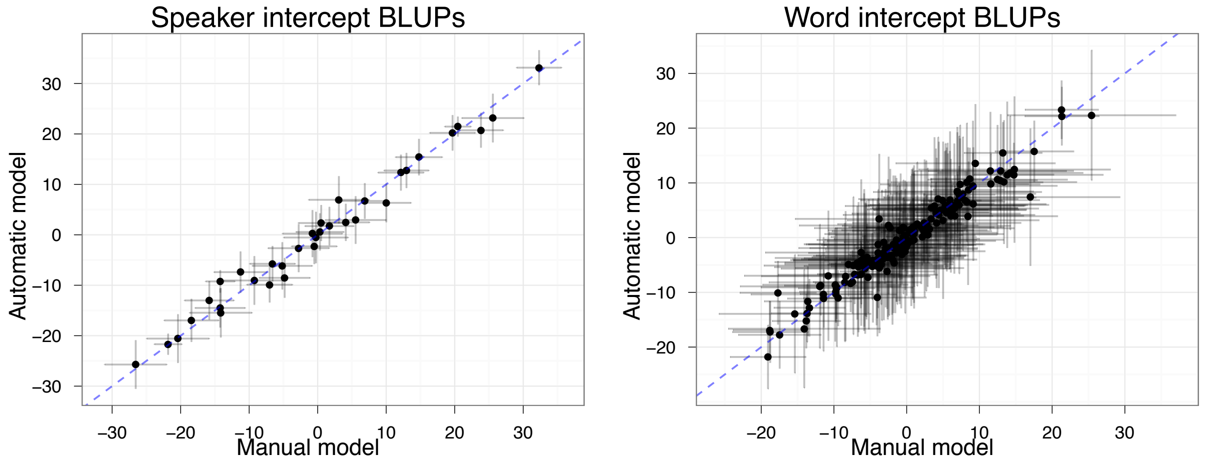

Figure 3.7: Estimated BLUPs for the by-speaker and by-word random intercepts. Horizontal and vertical positions correspond to values in the manual and automatic models, and horizontal and vertical errorbars corresponding to 95% confidence intervals in each model (1.96×standard error).

Thus, both the fixed effects and random effects are very similar for the two models. The similarity of the fixed effect coefficients means that the models make quantitatively similar predictions for the effects of place of articulation and first language. The similarity of the random effect BLUPs means that the two models predict similar deviations from the overall mean for each individual speaker and word.

Comparison of model predictions Comparing the automatic and manual models' predictions on the 25% of held-out data gives a sense of how similar their predictions are on unseen data. The two models make extremely similar predictions, differing by ≤ 5 msec for 90.2% of data points in the held-out set, and with correlation $r = 0.992$ and mean absolute difference 2.31 ms across all data points in this set. By comparison, measurements by two human transcribers on a subset of the PGWORDS dataset (see Sec. 3.5.3) differ by ≤ 5 ms for 90.0% of data points, and have correlation $r = 0.987$ and mean absolute difference 2.49 ms. Thus, the automatic and manual models make predictions that agree as well as two human transcribers.

3.9 Discussion

Summary We have described a machine learning approach to the problem of automatic VOT measurement which treats this task as a case of structured prediction. A function to measure positive VOT is learned from manual measurements, using a discriminative large-margin training procedure which aims to minimize the difference between predicted and actual VOT. The function takes as input feature maps which are specialized for the task of VOT measurement. In a first set of experiments, we showed that the algorithm achieves excellent performance for each of four datasets, when all data available for training is used, and in particular near-ITR performance on the three dataets were ITR data was available. In a second set of experiments, we showed that the algorithm is robust to decreasing the amount of training data, with performance remaining essentially constant down to about 50–250 training examples (depending on the dataset). Thus, the algorithm is adaptable to new datasets with relatively little effort. We also found that performance generally suffers for mismatched versus matched training and test corpora. Thus, the algorithm is learning something about the particular type of speech and measuring criteria used for each dataset.

Our algorithm generally outperforms previous work where automatic and manual VOT measurements are compared, with the caveat that precise comparison is difficult in some cases because of differences in the datasets and experimental setups used. There are several points of difference between our algorithm and previous work which may help explain our method's better performance. In contrast to the approaches of Stouten and van Hamme (2009), Hansen et al. (2010), and Yao (2009a), our system is trainable (rather than using a fixed set of rules), and thus can adapt to particular datasets and measurement criteria. In addition, the features used in our algorithm were specialized for the task of VOT measurement. This is a significant point of difference with Lin and Wang (2011) (along with our use of a different classifier and discriminative training), who also take a machine learning approach. Finally, our use of discriminative large-margin training allows us to train the measurement algorithm to minimize error on the particular quantity of interest, VOT.

We also evaluated our algorithm by comparing two mixed-effects regression models for the effect of several covariates on VOT in a dataset of laboratory speech: one model fitted using automatic measurements, and the other fitted using manual measurements. The two models were extremely similar, both in terms of fitted model parameters and predictions on held-out data. This shows that a study of how the covariates affect VOT in this dataset would have reached the same conclusions whether VOT measurements were done manually, or automatically using our method.

Future directions In this chapter, we have considered word-initial English voiceless stops, because we know that they are nearly always realized with a burst, and hence have positive VOT. (In TIMIT, for example, 99.6% of word-initial voiceless stops for words other than 'to' , which is sometimes flapped, are realized with a burst.) However, for stops which are not word-initial or not voiceless, this is often not the case. English stops in non-initial position are often not realized with a burst; for example, Randolph (1989) found that in 3 corpora of read speech (including TIMIT), 31% of stops occurring as syllable codas were realized with a burst, compared with 97% for syllable-initial stops. Voiced stops in English are sometimes realized with negative VOT, though estimates of how often such "prevoicing" occurs vary greatly (e.g., 23% in Lisker and Abramson, 1964, vs. 62% in Smith, 1978 for word-initial /b/ in isolated words; see Docherty 1992 for discussion) The negative VOT case is even more important for languages such as European French or Thai, where phonologically voiced stops are almost universally realized with prevoicing (Caramazza and Yeni-Komshian, 1974; Kessinger and Blumstein, 1997; Lisker and Abramson, 1964). Future work will deal with both the task of measuring negative VOT, and the task of deciding whether or not a burst occurred for a given stop. Both are necessary for our ultimate goal: an automatic measurement system that can take an arbitrary segment of speech and its orthographic transcription, and output a VOT measurement (positive, negative, or no burst) for each stop that is expected to occur.

The approach taken here combines knowledge about the cues human annotators use to measure VOT with machine learning techniques for predicting structured output, to tailor an algorithm to measure VOT nearly as accurately as humans, and which meets the three criteria laid out in the introduction: accuracy, trainability, and robustness. Given suitable features and training data, it would be straightforward to extend the approach taken here to other widely-measured phonetic variables where the output is a sequence of time points, such as vowel duration, segmenting a stop into different parts (closure, burst, frication), or the duration of vowel nasalization. More generally, for most phonetic quantities of interest (e.g., VOT, vowel formants, spectral measures for fricatives) measurement is a skilled task, and expert annotators usually use several types of cues (spectral, auditory, what the quantity "should" look like) in reaching a decision. The approach taken here is to supply these cues as features to an appropriate machine learning procedure to learn how to annotate a particular quantity. Knowledge about the annotation task is also tied to the algorithm's structure by using a specialized cost function related to the quantity being annotated, which is directly optimized using discriminative training. The results of the study reported in this chapter suggest that combining knowledge about the annotation task to be performed with appropriate machine learning techniques is a promising direction for designing algorithms to automate phonetic measurement in general.

CHAPTER 4

BIG BROTHER: BACKGROUND

In Chapter 1, we raised two questions about longitudinal variation: what are the dynamics of sound systems, and what are the sources of any observed dynamics? The reality television show Big Brother, where housemates live in an isolated house for three months, provides a fascinating opportunity to study these questions at the level of individual speakers in a controlled setting.

The remainder of this thesis is a study of synchronic variation and longitudinal variation in a corpus of spontaneous speech from the show. This chapter reviews previous work; the next describes the show, the corpus, and the datasets of phonetic measurements for five variables: voice onset time (VOT), coronal stop deletion (CSD), and formants for three vowels. In Chapters 6 and 7, we build models of synchronic and longitudinal variation for each variable.

In this chapter, we first briefly discuss previous studies that use reality TV data for linguistic research, or investigate phonetic and phonological variation in spontaneous speech corpora (Sec. 4.1–4.2). We then review previous work on longitudinal phonetic and phonological variation during adulthood (Sec. 4.3), and on factors conditioning variation in each of the five variables (Sec. 4.4).

4.1 Reality television as linguistic data

One strand of linguistic research using reality TV can be characterized as discourse analysis or pragmatics: data from reality TV are used to examine how language use contributes to meaning in social interaction, and how the situational context of a linguistic utterance affects its meaning. For example, Thornborrow and Morris (2004) study ‘gossip talk’ on a season of Big Brother, and argue that housemates talk strategically both to form alliances with other housemates, and to manage how they are seen by the viewing public. Wahl

(2010) examines African performances of California 'dude style' in *Big Brother Africa 3*. Poulios (2009) analyzes the construction of age identity in interaction on a Greek reality TV show, using conversational analysis. Other work makes use of the adversarial interactions common on reality TV, for example how impoliteness is expressed on *Dragon's Den* (Lorenzo-Dus, 2009), or how parents on *Honey We're Killing the Kids* respond to criticism from nutrition experts (Gordon, 2011). Finally, examples from reality TV can be used as part of a broader argument about pragmatics or discourse (e.g., Auer, 2005, 2010).

Studies in the second strand use data from a particular show to address linguistic questions that are well-aligned with the show's structure. For example, Sankoff (2004) uses data from the "Up" series of documentaries, in which the same British speakers are revisited every seven years from childhood on, to study longitudinal variation in individuals over the lifespan.[1] Bergmann (2006) uses speech from German *Big Brother* in a study of regional variation in German intonation. Here, *Big Brother* is a useful source of data because housemates come from diverse dialect regions. Cramer (2010) analyzes speech from characters on *Southern Belles: Louisville* as part of a study of the production of regional identity in Louisville, Kentucky. Data from the most prototypically-Louisville speaker is used as input for a focus group of Louisville residents on regional identity, taking advantage of the show's explicit focus on local identity.

Unsurprisingly, most studies of language on reality TV are in the first strand: reality TV is about performance and intense social interactions, making it a gold mine for studying topics like emotional speech, pragmatic meaning, or (im)politeness. The dramatic nature of reality TV also makes it suspect, a priori, as a convenient source of spontaneous speech data for linguistic research more generally. Because contestants are ultimately performing, their speech could be quite different from everyday conversational speech, raising similar concerns about ecological validity as when other types of media speech

[1] This study is discussed further in Sec. 4.3.1.2.

(such as broadcast speech) are used as linguistic data.[2] One contribution of the current study is to add another example showing that reality TV speech can be usefully used as a valid source of spontaneous speech data. There is no glaring evidence for important differences between our corpus and previous work on spontaneous speech for the variables we examine. Along with other studies in the second strand, the current study suggests that reality TV shows can be treated as "natural experiments" with ecological validity, which provide unique opportunities for asking linguistic questions related to the structure of the show, and which often would be difficult to ask in real life.

4.2 Phonetic and phonological variation in spontaneous speech corpora

Broadly speaking, empirical work on phonetics and phonological phenomena is characterized by two dimensions corresponding to the type of speech analyzed. The researcher can use speech elicited in a laboratory or other controlled setting, or in a non-laboratory setting; the speech can be either planned (isolated words, read passages) or spontaneous. These dimensions correspond to four types of speech data (laboratory/planned, etc.), all of which have been examined in some studies. The vast majority of empirical work on phonetics and phonology has used planned laboratory speech. A smaller strand has used corpora of spontaneous non-laboratory speech, such as the Buckeye Corpus (American English; Pitt et al., 2007), the Corpus of Spoken Dutch (Oostdijk, 2000), or a custom-built corpus (most sociolinguistic studies). These two categories, planned laboratory speech and spontaneous non-laboratory speech, are often what is meant when "laboratory" and "corpus" approaches to studying sound systems are contrasted.[3] While there are impor-

[2] See Tagliamonte and Roberts (2005) for a useful discussion of the benefits and risks of mining television shows for linguistic data.

[3] Note that the datasets used in "corpus" studies, such as the Big Brother corpus, do not necessarily satisfy the conditions (i.e., representativeness, balance) implied by the term in corpus linguistics.

tant exceptions that fall into the remaining two categories (laboratory/non-laboratory, spontaneous/planned), we adopt this shorthand as well.[4]

The pros and cons of laboratory and corpus studies are the same as for controlled experiments and observational studies in any field. There is a tradeoff between certainty about the conclusions of the study (what effect does a factor have, when all other factors are held constant?), versus their generalizability to other datasets. As Xu (2010: 334) puts it, the choice is "to maximize ecological validity or to maximize the level of control." The merits of each type of study for investigating phonetics and phonology have been extensively discussed (e.g., Beckman, 1997; Cole and Hasegawa-Johnson, 2012; Rischel, 1992; Xu, 2010). We believe that both approaches are ultimately essential for a full understanding of phonetic and phonological *variation* in particular. From this perspective, more studies of phonetic and phonological variation in spontaneous speech corpora are needed, simply because the majority of previous studies are laboratory experiments.

Previous work on variation in spontaneous speech corpora falls into two strands, roughly corresponding to corpus linguistics and sociolinguistics.[5] The main difference is whether the object of study is primarily phonetic and phonological phenomena in individual speakers, or language variation and change in speech communities.

The first strand consists of a small but growing body of studies, reviewed in part by Ernestus and Baayen (2011) and Cole and Hasegawa-Johnson (2012). For example, two of our variables have been examined in the Buckeye Corpus: coronal stop deletion and VOT (Raymond et al., 2006; Yao, 2009b). Some examples of questions addressed by this type of study are: to what extent do phonological phenomena traditionally described as categorical, like Dutch voicing assimilation or English nasal place assimilation, occur in

[4] A non-laboratory/planned example is TIMIT (see Sec. 3.4.1) A laboratory/spontaneous example is the HCRC Map Task Corpus of dialogues between speakers completing a shared task in a laboratory setting (Anderson et al., 1991).

[5] Kendall (2011) provides a very useful discussion of the relationship between corpus linguistics and sociolinguistics, and discusses other work addressing the two. However, phonetic and phonological variation are mostly not discussed, since corpus linguistics as a field has tended to focus on text rather than speech corpora.

spontaneous speech (Dilley and Pitt, 2007; Ernestus et al., 2006)? What are the factors affecting how frequently segments are reduced in continuous speech, and what do observed patterns of segmental reduction imply about speech production (Bell et al., 2009)? The questions addressed typically focus on the structure of aspects of speakers' phonetics or phonology, and less on how phonetic and phonological variables vary across speech communities, or change over time.

The second, much larger strand is sociolinguistics, which beginning with Labov's pioneering study on Martha's Vineyard (Labov, 1963) has primarily examined phonetic and phonological variation. Sociolinguistic studies analyze collections of spontaneous speech, elicited in sociolinguistic interviews, tape-recorded conversations between friends, and other settings. These collections are spontaneous speech corpora, though with significant differences from the databases which are usually understood by the term: they tend not to be publicly available and are designed for sociolinguistic investigation of a particular community, rather than speech and language technology applications. A typical corpus includes a stratified sample of speakers with respect to various social categories (e.g., sex, class, or locally-important variables). Studies in this strand focus on variation in particular phonological variables from the perspective of understanding variation and change in speech communities, and less on speakers' phonetic or phonological knowledge about the variable per se. For example, a study of fronting of the GOOSE vowel in a particular English-speaking community (see Sec. 4.4.3.1) might address what phonological environments promote fronting, which social groups front more often than others, and whether the structured variation present in the community is stable or part of an ongoing change.

A growing body of work, which largely falls under the label 'sociophonetics', combines these strands: a particular phonetic or phonological variable is investigated both in its own right, and also to address variation and change in that variable in a given community, using a corpus which allows both types of investigation (e.g., Hay and Sudbury, 2005; Keune et al., 2005; Raymond et al., 2006). The current study falls into this category.

For VOT and CSD in particular, we are both able to contribute to a substantial existing literature (of laboratory studies for VOT, and sociolinguistic studies for CSD) on the structure of synchronic variation, and examine the dynamics of each variable in the house.

4.3 Longitudinal variation in adulthood

Recall the questions to be addressed by building dynamic models for the Big Brother data: what time dependence do variables show within individual housemates, and what are the sources of any observed dynamics? The literature on longitudinal variation in individuals more generally has addressed three closely related questions:

1. What are the dynamics of phonetic and phonological variables in individuals during adulthood?
2. How and why do individuals differ in the dynamics of particular variables?
3. How and why do variables differ in their dynamics within individuals?

Two types of previous work address these questions: *long-term studies* examine individuals on a timescale of years, and *short-term studies* on a timescale of minutes to days. We review long-term and short-term studies of phonetic and phonological variables, then discuss proposals for the link between short-term and long-term change in individuals, and how the current *medium-term* study addresses questions raised by the short-term and long-term literatures.

4.3.1 Long-term studies

Individuals acquire their first language during childhood, and continue to refine aspects of their linguistic system through adolescence. The traditional view is that an individual's linguistic knowledge is largely fixed following adolescence.[6] This assumption underlies

[6] For example, from a standard textbook: "For the stages of life beyond young adulthood, our best evidence indicates that once the features of the sociolect are established in the speech of young adults,

the famous *apparent time* construct—that a variable's diachronic development in a speech community can be read off from its distribution among adults of different ages—which has been widely used since its introduction by Labov (1963, 1966) to study changes in progress (Bailey et al., 1991). However, it has always been known that post-adolescent stability is only an approximation, and a significant body of studies addresses the plasticity of speakers' linguistic systems during adulthood. This literature can be divided into two broad categories, depending on whether the individuals under study have remained in the same speech community (*panel studies*), or have moved between speech communities (*dialect change studies*).[7]

In the first type of study, a fixed group of speakers in a speech community (the panel) is studied at several time points. Panel studies are complementary to trend studies, where a different group of speakers is studied at each time point. Trend and panel studies together are sometimes called *real-time* studies in sociolinguistics (in contrast to apparent-time). The two types of studies are reviewed by Sankoff (2005, 2006, 2012).[8] In studies of dialect change, acquisition, or shift, a group of individuals is studied who have moved between speech communities where different dialects are spoken. We call this the "dialect change literature" for simplicity; work in English is reviewed by Siegel (2010).

For logistical reasons, it is very difficult to study individuals at several points in time.

under normal circumstances those features remain relatively stable for the rest of their lives. Even when linguistic changes take root in the speech of younger people in the same community, the older people usually remain impervious to it, or nearly so" (Chambers, 2003: 197).

[7] We will not discuss a third relevant literature, on "vocal aging": changes in speakers' phonetic parameters due to the physiological effects of aging (see e.g., references in Reubold et al., 2010). Though potentially very relevant when considering change in individuals over many years (e.g., Harrington et al.'s study of Queen Elizabeth II over 50 years, discussed below), no vocal aging effects are expected on the timescale of Big Brother.

[8] Note that Sankoff does not explicitly state that speakers in a panel study must remain within the same community, and "panel studies" could be taken more broadly to include studies of mobile speakers as well. However, the studies discussed by Sankoff (and in the real-time sociolinguistic literature more generally) largely consider speakers who remain in the same speech community, since real-time studies in this literature are of interest primarily for studying variation and change in a given speech community, the questions which they can address being complementary to those which can be asked in apparent-time studies. We use "panel studies" to refer to studies of speakers who stay in the same speech community, for convenience.

Accordingly, there are many more trend studies than panel studies, and most studies in the dialect change literature consider speakers at a single time point in a second-dialect (D2) environment and infer change relative to their native dialect (D1). Nonetheless, a sizable literature has developed on change in different aspects of individuals' linguistic systems over time. We restrict our review to studies that are most relevant to understanding the dynamics of phonetic variables within speakers on Big Brother, all of whom are past adolescence. With a few exceptions of unusual relevance (e.g., a study of accent change in British university students), we only consider studies of phonetic or phonological variables, in post-adolescent speakers, where speakers are measured at several points in time.

4.3.1.1 Panel studies

Montreal The most extensive panel study to date followed 62 Montreal French speakers over 24 years. 120 speakers were recorded in 1973, stratified by age, sex, and class (Sankoff and Sankoff, 1973). In 1984, sixty of these speakers were re-recorded, along with 12 younger speakers (Thibault et al., 1990). In 1995, 12 of the original speakers and two of the speakers added in 1984 were re-recorded (Vincent et al., 1995). The dynamics of many syntactic, morphological, and lexical variables have been examined for individuals in this panel (Blondeau, 2001, 2006; Blondeau et al., 2002; Daveluy, 1988; Dubois, 1992; Labov and Auger, 1993; Lessard, 1989; Thibault, 1991; Thibault and Daveluy, 1989; Wagner and Sankoff, 2011), in addition to two types of phonological variable.

Two studies examine whether individuals changed their realizations (formant frequencies) of vowels that are changing at the community level. Yaeger-Dror (1994) shows that some change occurs for most speakers, but it is not clear in which vowels change occurs, or whether it is in the direction of community change or not (due to how results are reported). MacKenzie and Sankoff (2010) find a mixture of stability and change for particular vowels for particular speakers, but the changes observed in individuals often do not line up with the direction of change in the community; they conclude that their vexing results are likely due to small sample size for most vowel/speaker pairs.

A clearer picture emerges for a second variable, the realization of /r/ as apical [r] or dorsal [R], which has been examined in several trend and panel studies (Blondeau et al., 2002; Sankoff and Blondeau, 2007, 2010; Sankoff et al., 2001). The pronunciation of /r/ began shifting from [r] to [R] in Montreal between 1950 and 1970, at the community level.[9] A central question addressed in these studies is whether and how individuals change, in the midst of this community change.

Sankoff and Blondeau (2007) combine trend and panel components to compare the realization of [R] in 1973 and 1984, both within individuals and at the community level. They examine data from a panel of 32 speakers recorded at both time points and stratified by class, age, and gender, as well as two groups of 32 speakers, one from each time point, with demographics matched to the panel as closely as possible. Community use of [R] increased from 63.8% to 77.8% from 1973 to 1984, continuing the long-standing trend.

At the level of individuals, the picture was more complicated. Each speaker at a given time point is classified as an [r] speaker (<17% [R]), an [R] speaker (>84% [R]), or a variable speaker.[10] Of 12 speakers who were [r] speakers in 1973, 10 showed no change by 1984, while two speakers became variable. All 10 speakers who were [R] speakers in 1973 remained [R] speakers in 1984. Of the remaining 10 speakers who were variable in 1973, three showed no significant change in their rate of [R] usage by 1984, and seven became [R] speakers. As is often observed in cases of change in progress, women were leading: of the 10 most advanced speakers ([R] users in both years), eight were women; of the 10 least advanced speakers ([r] users in both years), seven were men. However, there were no clear gender differences in which speakers were variable, or which speakers changed between the two timepoints. The overall picture is of change in the community, but rel-

[9] A number of other surface realizations of /r/ also occur. [r] and [R] are the most common, and Sankoff et al. focus on variation between them.

[10] It is not clear why this range in particular is chosen.

ative stability in the majority of individuals, with a minority showing significant change in the direction of the community change. Variable users seem to be particularly susceptible to change: only 16% of [r] users in 1973 became variable users by 1984, while 70% of variable users in 1973 became [R] users by 1984.

Queen Elizabeth II The most extensive longitudinal studies of an individual (panel of one speaker) have been conducted by Harrington and colleagues on Christmas Day broadcasts given by Queen Elizabeth II between 1952 and 2002, examining change in vowel formants and fundamental frequency. As "perhaps the only acoustically high-quality annual recordings of speech data from the same person over a fifty-year period," these broadcasts present a remarkable opportunity to study the plasticity of an individual's speech during adulthood (Harrington, 2007: 127).

The formants of many of the queen's vowels have changed significantly over this time period, both for monophthongs and diphthongs (Harrington et al., 2000a,b, 2005). The changes have generally had the effect of moving the queen away from a traditional upper-class Received Pronunciation (RP) accent, and towards a "more modern, less aristocratic form of RP" (Harrington, 2007: 128) Most strikingly, for several vowels which are examined in detail, changes in the queen's pronunciation parallel change occurring at the community level in RP (Harrington, 2006, 2007).[11] The queen's vowel space has also expanded in F1, with low vowels lowering and high vowels raising. Some (though far from all) of the change observed in vowel qualities can be attributed to vocal aging: pitch and F1 (across all vowels) decrease substantially over the fifty-year period, a pattern generally consistent with the vocal aging literature (Harrington et al., 2007; Reubold et al., 2010).

Scandinavia In their large real-time study of pronunciation in Copenhagen, Brink and Lund (1975) "examined recordings of several Danish speakers over many decades, in one case with a 50-year interval, and found all speakers' phonologies to be extremely stable"

[11] Fronting of GOOSE, lowering and backing of TRAP, and tensing of final /ɪ/ in words such as "happy."

(Sankoff, 2005: 1004).[12] In a second Danish panel study, Gregersen et al. (2009) examine the short (æ) variable, by 43 speakers in Copenhagen and Næstved, a nearby town. The variable can be realized as [æ] (standard) or [ɛ] (socially and regionally marked). Three quarters of speakers show no change between the two timepoints, and of the remaining quarter, some show change in each direction (lower or higher %[ɛ]).

Nahkola and Saanilahti (2004) examine recordings from 1986 and 1996 of 24 speakers in a Finnish town (Virrat), and consider 14 phonological, morphological, and morphophonological variables. Their broad conclusion is that if a variable is acquired as categorical or near-categorical, it is unlikely to change over the lifespan; if it is acquired as variable, change is possible.[13] With regard to the stability of variables within the individual, 39% of all variables within a given speaker change by at least 10% between 1986 and 1996, and in only a few cases does a variable's usage change "radically" between the two years. Nahkola and Saanilahti acknowledge the arbitrariness of their 10% cutoff, and raise an important point: it is not currently possible to say how much of a change between the two corpora represents a "truly relevant" change, because we do not know how much an individual's usage of a variable is expected to fluctuate from day to day (p. 90). This observation is a key motivation for studying *trajectories* of variation in Big Brother, and we return to it below.

As part of a combined trend and panel study of change in the consonantal variable (d) in another Finnish town (Hanhijoki), Kurki (2003) examines (d)'s usage by 11 speakers at two time points 10 years apart. Nine speakers do not change, while two show "rather great" change in the opposite direction of the community change. However, no significance tests are given.

Other studies Bowie (2005) examines 13 phonological variables from religious addresses

[12] Unfortunately this work is in Danish, so I am unable to expand on Sankoff's summary.

[13] Because all results are reported in terms of groups of speakers born in similar years, it is not possible to say much more about change within individuals.

by five church elders in Utah over several decades: four speakers at two timepoints 20 years apart, and one speaker at three timepoints over 40 years. For 10 variables, no significant change is found for any speakers. For three variables, some speakers show significant changes, but differ in the direction of change. Bowie concludes there is some support for pronunciation change during adulthood.[14]

As part of a larger trend study of change in Received Pronunciation, Bauer (1985) briefly examines productions of three vowels by the same three RP speakers, at two time points approximately 20 years apart, reading a short passage. For all speakers F1 and F2 have changed for each vowel, in some cases in the direction of ongoing community change. However, no tests of significance are given, and Bauer notes that vocal aging may be responsible for some of the observed change (overall lowering of F1).

4.3.1.2 Dialect change studies

Saxon migrants A natural setting for studying dialect change is internal migration, where groups of speakers move from one dialect region to another in a single country, often in search of work. A remarkable study followed 56 speakers (aged 12–52) of Upper Saxonian Vernacular (USV), a highly stigmatized German dialect spoken in Saxony (former East Germany), who moved to two locations in the former West Germany following German reunification (Barden and Großkopf, 1998). Speakers were recorded eight times over two years. Reporting on a subset of this study, Auer et al. (1998) examine 14 phonological variables which are characteristic of USV at three time points, focusing on how much speakers decrease their use of the USV variant relative to standard variants (from Standard German, or the dialect of their new location). They test whether five "objective" and five "subjective" criteria for a variable's salience (see Sec. 4.3.1.3) predict the relative

[14] Relatively few tokens per variable (30 or fewer) are considered at each timepoint, raising the possibility that some changes which did occur may have gone undetected (i.e., Type II error: see Sec. 4.3.1.3).

amount of loss of different USV variables.[15]

The overall pattern is decreased use of USV forms, for nearly all variables. Salience turns out to be helpful only for continuous variables (and not dichotomous variables), which can be realized as intermediate (between USV and the standard) or strong (USV) forms. Salience predicts the amount of decrease of intermediate forms only. Both subjective and objective criteria matter, but subjective criteria dominate.

The results are not broken down by individuals, so we cannot say whether some speakers are much more stable than others (as observed for Montreal /r/). However, the speaker's social network and attitude play an important role. Speakers who formed no strong social ties in West Germany and whose attitude was oriented towards Saxony (and against West Germany) behaved differently from other migrants. The former actually increased their usage of USV forms, bucking the general trend, and mirroring their attitude towards their situation (Barden and Großkopf, 1998, summarized in Auer and Hinskens, 2005).

English university students University campuses, where students from many different dialect regions come into contact, form another natural setting for studying dialect change. Evans and Iverson (2007) examine the dynamics of vowel perception and production for 23 students (aged 17–18), whose home dialects were a variety of Northern English (NEng). Speakers were recorded at four time points over two years, beginning before their first year of university. Of interest was how the northern students' vowels shifted in response to contact with the prestige dialect, Standard Southern British English (SSBE), which was expected to be the majority dialect at their university.

We discuss only the production results. At each timepoint, speakers produced words corresponding to 10 lexical sets ("beat", "boot") in carrier sentences. The words contained

[15] For example: variables involving a phonemic distinction are more salient than those involving subphonemic differences (objective); the more aware lay speakers are of a variable, the more salient it is (subjective).

eight distinct vowels in NEng, versus nine in SSBE, due to the FOOT/STRUT merger in NEng (see Sec. 4.4.3.1). F1, F2, and duration were analyzed for FOOT, STRUT, and BATH, to test for change in these vowels' qualities over two years. These three vowels correspond to the most socially-salient differences between NEng and SSBE: the merger of FOOT and STRUT, and the pronunciation of the BATH lexical set.[16] Speakers changed F1 and F2 of FOOT and STRUT over time towards a more centralized pronunciation. One interpretation is that they maintain the merger between these lexical sets, but the realization of the merged set shifts, from a point similar to SSBE FOOT to intermediate between SSBE FOOT and STRUT. Speakers also change F1 and F2 of BATH over time, towards PALM, as expected if they are shifting towards the SSBE pattern (BATH=PALM). However, there is no change in duration, even though BATH is significantly longer in SSBE than in NEng.

The other vowels are not explicitly tested for change, but visually seem stable at the group level.[17] This includes several vowels whose realization varies significantly among British dialects (such as GOOSE; see Sec. 4.4.3.1), suggesting that speakers are not simply moving their entire vowel space towards SSBE. Individual results are not explicitly reported, but some differences between subjects are noted: the realization of FOOT/STRUT changed for "almost all" subjects, while "some" changed their realization of BATH.

Brazilian/English bilingual A pioneering study by Sancier and Fowler (1997) does not concern dialect change per se, but the effect of a bilingual speaker's ambient language on her speech production in both languages. A 27 year-old Portuguese/English bilingual was recorded reading a list of sentences in each language, containing words beginning with initial voiceless stops, at three time points: during a trip to Brazil, and in the US before and after the trip. In each case, she had been in the country for several months prior to recording. Overall, the VOT of her Portuguese stops are shorter than for her English

[16] For northern speakers BATH is pronounced like TRAP, and for southern speakers like PALM.

[17] The other vowels are not considered "to avoid multiple statistical tests" (3818), given that the authors' previous work suggests that only FOOT, STRUT, and BATH are likely to change (Evans and Iverson, 2004).

stops, whether she is in the US or Brazil, as expected given the different timing patterns of stops in these languages (see Sec. 3.8). However, she produces stops in both languages with shorter VOTs in Brazil (where the ambient language has shorter VOTs) than in the US (where the ambient language has longer VOTs). Sancier and Fowler conclude that the speaker shows "gestural drift": the timing of her articulatory gestures change such that her VOTs shift towards the ambient language of the community.

North Americans abroad In one of the largest studies of second dialect acquisition, Foreman (2003) examines 34 Americans who moved to Australia at different ages, and have lived there for different amounts of time. Six phonological variables are considered which differ substantially between Australian English and American English. Longitudinal data is available for six speakers. Four speakers, all of whom had lived in Australia since the 1970s, show no significant changes for the six variables between recordings taken in 1989 and 1999. For two other speakers, five recordings are available from 1974–2001, beginning six months after their arrival in Australia. After 27 years, the speakers use 2.3% and 20.6% Australian variants, pooling across all variables. There is significant variation in how much different variables change, for both speakers.

In an early study of dialect change, Shockey (1984) examines flapping of intervocalic stops by four Americans who have moved to England. She examines herself at two time points, six months and 3.5 years after arriving in England. After six months she still uses the flap allophone categorically in flapping environments; after 3.5 years she uses it variably, but still more frequently than the other three speakers, who have lived in England 8–27 years.

A number of other dialect change studies examine Americans and Canadians who have moved to other English-speaking countries, but are outside the scope of this review (phonetic and phonological variables, post-adolescents, multiple recording times) (Chambers, 1992; Munro et al., 1999; Tagliamonte and Molfenter, 2007; Trudgill, 1986).

Media In a well-known early study, Trudgill (1983) examines the use of two consonantal variables in songs by the Beatles and the Rolling Stones between 1963 and 1969. Each variable has British and American variants. There is a steady decrease in the use of American variants by both bands during this period, which is attributed to the rising dominance of British artists in rock music. Unfortunately (for our purposes), the results are not broken down by individual singers. Also, it is not clear whether Trudgill views the decreases in American forms as dialect change (which reflects the singers' accents outside of performances), or style shifting.

In a remarkable lesser-known study, Prince (1987, 1988) examines over 10000 tokens of five vocalic variables in performances by the Yiddish singer Sarah Gorby over 40 years. Each variable differs between Gorby's stigmatized native dialect, Bessarabian Yiddish (BesY), and the prestige dialect, Standard Yiddish (StY). The tokens are grouped into three time periods, and logistic regression analyses are performed to assess whether Gorby's use of each variable changed. Gorby shows a mixture of stability and change. About one-half of tokens are discarded at the outset because they are from words where *only* the StY variant is used, suggesting stability.[18] Two of the variables show no change. In at least two of the other three variables, the use of StY forms increases markedly over time. The variables' dynamics are affected by characteristics of the host morpheme: open-class words change less than closed-class words (Prince attributes this to less attention paid during production of closed-class words); words with different parts of speech and free versus bound morphemes also have different dynamics. Prince argues the mechanism of change is long-term accommodation to StY, and that some of the huge differences in dynamics between variables can be attributed to social factors, such as differing levels of stigmatization.

Sankoff (2004) examines two subjects of Michael Apted's *Up* series of documentaries,

[18] For some words (which are included in the analysis) the BesY variant is categorically used, but how many is not indicated.

which has revisited a group of 14 British children every seven years from age seven on. Both speakers are from Northern England, and each eventually moves between several dialect regions and experiences extraordinary social mobility (one upwards, the other downwards). Sankoff examines their productions from ages 7–35 (five timepoints) of BATH and STRUT, the two vowels found by Evans and Iverson to change for Northern English students after beginning university. Both speakers show "some significant phonetic, and possibly phonemic, alternations to their speech after adolescence" (136); Sankoff views their malleability as exceptional, possibly due to their highly unusual personal histories.

4.3.1.3 Discussion

The long-term literature sheds significant light on the three questions raised above.

Stability and change What are the dynamics of phonetic and phonological variables in individuals during adulthood? Recall that the traditional assumption is that there is no change during adulthood. The studies reviewed above suggest that adults who stay within a speech community, stability is the default, with a minority showing significant change, often in the direction of community change or the standard language. This picture generally holds for change within the individual for morphological and syntactic variables as well (Sankoff, 2012). The picture is similarly mixed for dialect change studies: when individual results have been reported, speakers show highly variable amounts of change for the same variable. For both phonetic/phonological and other types of variables, huge interspeaker variability is the norm (Siegel, 2010: 51).

The prevalence of stability for phonetic/phonological variables in the studies reviewed is striking. In every panel study where individual results are reported for more than three speakers, the majority of cases show no change: 66% of speakers for Montreal /r/, 75% of speakers for Danish (ae), 61% of cases in Virrat, 88% of cases in Utah.

There is an important distinction between stability and change in interpreting the results of longitudinal studies. It is not surprising to find that an individual's use of a variable has changed between two recordings taken years apart, a priori. When the recordings are of spontaneous speech (as is usually the case), the variable may be used at different rates in the two recordings because of differences in speaking style, the particular lexical items used, or many other factors which cannot be controlled for, given the limited amount of speech available from each time point. Differences may result from the fact that the two recordings are conducted with different methodologies or by different fieldworkers (Bailey and Tillery, 1999; Trudgill, 1988). There is also the possibility raised by Nahkola and Saanilahti that speakers vary randomly from day to day, in which case an observed difference between two time points may simply be noise. Given all the possible reasons for observing change, what is surprising is *not* observing change.[19]

However, there is a crucial caveat to interpreting findings of stability, which is that they have never (to our knowledge) been subjected to statistical tests to assess Type II error: concluding there was no change when in fact change occurred.[20] Type II error depends on a number of factors, most importantly (for our purposes) sample size, and is assessed by a power analysis. A worrying example suggests that Type II error (β) may be a significant problem in long-term studies. We calculated power ($1 - \beta$) for the observed difference in proportion of [R] usage between 1971 and 1984 for the three "stably variable" speakers in the Montreal study (those for whom Type I error $\alpha > 0.05$, in Sankoff and Blondeau, 2007), by applying a two-sided sample proportion test to the data in Sankoff and Blondeau's Table 12. Speakers mostly had between 100 and 120 tokens per

[19] This is the reverse of the situation in experimental studies, where a null result cannot be easily interpreted (because it has many possible causes), while a positive result is surprising. Because findings of stability are surprising, we disagree with Bowie (2005: 50), who finds them "not terribly interesting... a finding, but a finding of a lack of a finding."

[20] More precisely, concluding that the null hypothesis of no change cannot be rejected, when in fact it is false. In a closely related context, Hinskens (1996: 386) discusses Type II error in an apparent-time study of regional dialect leveling: the risk of concluding a feature is not leveling when in fact it is (but the effect is too small to detect).

timepoint (p. 568); to be conservative (higher power), we assumed 120 uniformly. Power for the three speakers was 0.12, 0.19, and 0.28, implying a Type II error of 0.72–0.88, far higher than standard cutoffs (0.2–0.25). Thus, if a given speaker *had* changed their [R] usage by the empirically observed amount, there would be a 72–88% chance of concluding they had not—which is what was concluded. Thus, for these speakers we cannot rule out the possibility that they did not change (because $\alpha > 0.05$) or the possibility that they did change (because $\beta > 0.2$). This example suggests power analyses are crucial for future long-term studies—which generally have smaller sample sizes per timepoint than the Montreal study—and that existing findings of stability merit further scrutiny. However, for our purposes we will continue to interpret the long-term literature as showing a remarkable degree of stability, with the caveat that some observed stability may in fact be changes which were too small to detect.

Despite the prevalence of stability, it is clear that some speakers show considerable change during adulthood. Most dialect change studies show many individuals changing at least to some extent towards the D2 variant in each variable studied (Siegel, 2010: Ch. 2).[21] For panel studies, the clearest cases are those with a convincing interpretation for the observed pattern of change (since these are least likely to be due to noise): Queen Elizabeth II's vowel space has undergone extraordinary change in adulthood, largely towards middle-class RP; Sancier & Fowler's bilingual produces stops with VOTs shifted towards the norm in her ambient language; Montreal French speakers who change their use of /r/ do so almost exclusively in the direction of community change; Sarah Gorby shifted largely towards increased use of standard Yiddish forms.

Interspeaker variation Relatively few panel studies have discussed causes for interspeaker differences in how much a variable changes, probably because the small number

[21] A notable exception is Stanford's studies of dialects spoken by different Sui clans, where almost *no* change is observed for any variable for women who marry into different clans, apparently due to an extremely strong link between dialect features and clan identity (Stanford, 2007, 2008).

of speakers in most studies makes generalization difficult. Naro and Scherre (2003) and Wagner (2008) found that speakers with more education (during the study period) were more likely to change, for phonological and morphological variables where one variant was part of the standard language and was taught in schools. In Montreal, Blondeau et al. (2002) argue that increased use of [R] reflects upward social mobility, and Wagner and Sankoff (2011) find that upper-class speakers increase their use of the inflected future despite ongoing community change away from this form, perhaps because it has become prestigious. Educational level and social mobility could make change more likely either because they entail more exposure to the standard variant, or because the speaker comes to identify more with standard variant users.[22]

There has been much more discussion of sources of individual differences in dialect change. Siegel (2010: Chapters 4–5) summarizes the evidence for effects of age of acquisition (when the speaker moved to the D2 area), length of residence, gender, occupation, amount of interaction with D2 speakers, attitudinal factors (including motivation to learn the D2), and social identity.[23] He concludes that the most important factors are age of acquisition (the expected amount of change decreases significantly following adolescence); the amount of interaction with D2 speakers, and possibly how much the speaker identifies with the D2 community (p. 120).

Intervariable variation The most important factor conditioning whether a variable is likely to change over an individual's lifespan is what level of linguistic structure it reflects. There is consensus that both for static and mobile speakers, lexical or syntactic variables are more susceptible to change than phonological and phonetic variables (Labov, 1994: 84; Sankoff, 2005; Siegel, 2010: 92).[24]

[22] We discuss this type of ambiguity further in Sec. 4.3.3.1.

[23] "... the part of a person's self-image based on the characteristics and attitudes of the social group or groups which that person belongs to or aspires to belong to" (106).

[24] Morphological features are somewhere in between, but "more lexical" aspects of morphology may be more malleable (Sankoff, 2005).

Especially in dialect change studies, another important predictor of a variable's susceptibility to change is rule complexity: how complex an adjustment the D1 learner would need to make to her existing sound system to acquire the D2 variant. For phonetic and phonological variables in particular, purely phonetic adjustments should be easiest to acquire (D1 and D2 share a phoneme, but pronounce it differently), predictable phonological alternations harder, and arbitrarily phonemic splits very hard. This basic idea has been applied in many dialect change and acquisition settings to help explain which variables are more likely to be acquired, by both children and adults (e.g., Chambers, 1992; Payne, 1976, 1980; Wells, 1973, 1982). The role of rule complexity may differ for features that are being lost versus gained: complex rules may be hardest to acquire, but easiest to lose (Auer et al., 1998: 168).

Salience is widely invoked in the dialect change literature: the more salient a variable, the more likely it is to be gained (in dialect acquisition) or lost (in dialect shift). The concept of salience dates back to Schirmunski (1929, 1930), who applied it (*Auffälligkeit*) at the level of whole dialects (i.e., community-level); more recently it has been invoked to explain the likelihood of individuals taking up or losing a particular variable (e.g., Auer et al., 1998; Kerswill and Williams, 2002; Trudgill, 1986, and references therein). Intuitively, variables that are more noticeable or prominent will be more likely to change. This hypothesis seems reasonable, since you must be able to notice a variable to change it.[25] However, exactly what makes a variable salient is highly controversial. For the phonetic/phonological variables we will consider in Big Brother, three widely-accepted criteria for salience are relevant:

- *Greater phonetic distance*: Variables with a greater phonetic distance between the old and new variants are more salient.

- *Phonemicity*: Variables involved in distinguishing meaning are more salient than

[25] This holds for acquiring features of a second language as well, as pointed out by Siegel (2010: 127).

those which are not.

- *Social salience*: Laypeople attach meaning to the variable: they notice and comment on it, it shows up in imitations of the dialect or in lay dialect writing, laypeople associate it with particular demographic or regional groups, etc. Variables with greater social salience are more susceptible to change.

Accounts of salience differ on the non-trivial questions of what phonemic distance and phonemicty actually *are*, which we abstract away from here.

4.3.2 Short-term studies

Short-term studies examine a very general phenomenon: aspects of one's speech often shift under exposure to other people's speech, both in conversations and in laboratory settings. The direction and magnitude of these shifts are strongly mediated by social and situational factors, such as gender, social dominance, and attitude towards the interlocutor. We consider only shifts in the sound system, but such effects are widespread at higher levels of representation as well (Pickering and Ferreira, 2008; Pickering and Garrod, 2004).

Individuals have been found to shift their speech along many dimensions. Most studies have considered paralinguistic parameters, such as vocal intensity, pitch, speaking rate, or pause duration. This substantial literature is partially reviewed, from a phonetics perspective, by Nielsen (2008) and Babel (2009). Below, we focus on the growing body of work examining shifts in phonetic and phonological parameters, with some discussion of work from the broader literature to provide context.

4.3.2.1 Preliminaries

Causes of short-term shifts The key theoretical issue in the short-term literature is *why* individuals shift, on which there are two broad viewpoints. One is represented by Communication Accommodation Theory (CAT), which considers shifts in both speech and non-verbal communication during social interaction (Giles et al., 1991; Shepard et al., 2001).[26] CAT proposes that individuals use language to manage social distance from their interlocutor, by one of several *accommodation* strategies: *convergence*, for example to gain approval or express solidarity; *divergence*, for example to express disapproval or emphasize social differences; or *maintenance*, for example maintaining the same speech style despite accommodation by the interlocutor. The other viewpoint is that the shifts are largely automatic and unconscious, the reflection of a broad human tendency towards synchrony of actions with perception via the "perception-behavior expressway" (Bargh and Chartrand, 1999; Chartrand and Bargh, 1999; Dijksterhuis and Bargh, 2001). However, the link between perception and action is mediated by social factors, such as those invoked by CAT, that influence the direction and degree of shift.

We call these the *social* and the *automatic* positions. The debate will turn out to be largely unimportant for our study of phonetic dynamics in Big Brother, but frames much of the work reviewed below.

Terminology There is significant confusion in the literature on the meaning of terms referring to short-term shifts in speech: accommodation, convergence, divergence, and imitation.[27] 'Accommodation', 'convergence', and 'divergence' are traditionally used within the SAT/CAT framework: accommodation is the subjective sense of adjusting how one communicates in interaction, while convergence and divergence refer to actual shifts on some dimension. Thus, these terms are associated with the social position. 'Imitation' suggests automaticity, and is traditionally used by authors taking an automatic position, such as Bargh and colleagues. However, 'spontaneous imitation' refers to an influential experimental paradigm introduced by Goldinger (1998) for eliciting short-term shifts (described below). Although Goldinger was firmly in the automatic camp (hence 'imitation'),

[26] Developed from the earlier Speech Accommodation Theory, which considered speech only (Giles and Powesland, 1975).

[27] Even more terms exist (alignment, entrainment, resonance), but are used less frequently.

the term has come to be used more generally. Finally, the terms 'phonetic convergence' and 'phonetic imitation' are often used to refer to the general phenomenon of short-term shifts, particularly in phonetic/phonological variables such as VOT or vowel formants. It is not surprising given such overlap that the different terms are increasingly used interchangeably. (e.g., Abrego-Collier et al., 2011; Babel, 2011). Nonetheless, the terms retain strong connotations on the sources of observed short-term shifts, and it is worth clarifying how we will use them.

'Shift' is preferred, as neutral on the causes of an observed short-term change. 'Converge(nce)' and 'diverge(nce)' are used as shorthand for "shift towards/away from the interlocutor", with no implication of social motivations. 'Accommodation' is used only when the author invokes social motivations. 'Imitation' and 'spontaneous imitation' are used only to refer to Goldinger-like imitation paradigms, with no implication of automaticity. Rather than 'phonetic imitation' or 'phonetic convergence', 'phonetic shift' is used as shorthand for "shifts in phonetic or phonological parameters during short-term interaction." Note that this terminological clarity comes at the expense of often using different terms when describing the results of a study than are used by the study's authors.

4.3.2.2 Shifts in general

In short-term studies, shifts in an individual's speech can be measured in two ways: along some acoustic dimension(s), or using a perceptual measure of overall similarity. Most studies in the broader short-term literature have used acoustic measures. For example, Natale (1975) tested whether the mean vocal intensity of subjects' speech shifted towards the intensity of an interlocutor in two types of interaction: dialogue with an interviewer (21 subjects), and unstructured conversation with another subject (50 same-sex dyads). In the first setting every subject's vocal intensity shifted towards the interviewers; in the second setting conversation partners converged in vocal intensity, but individual results are not reported. More convergence obtained for subjects who scored higher on a social desirability scale.

The perceptual measure was first used in a seminal study by Goldinger (1998). Simplifying somewhat, eight subjects were recorded reading a list of words, then heard these words read by a set of model talkers, then performed a shadowing task (repetition of a speech stimulus as fast as possible) on these words read by the same talkers.[28] How similar a subject's shadowed production of a word was to its target production (by the model talker), relative to its baseline production, was assessed by an *AXB task*: a panel of (different) subjects heard the baseline (A), target (X), and shadowed (B) productions, and decided whether A or B was "a better imitation of" X. The percentage of votes for B is a gestalt perceptual measure of its similarity to X, which we call *AXB similarity*; a score of at least 50% indicates the subject shifted the word's pronunciation towards the target production. By this measure, subjects consistently shifted towards the target productions. Because subjects had no reason to consciously shift towards the target (no instructions to do so, or social interaction), Goldinger called this behavior *spontaneous imitation*. Later work extended the timescale of this effect (Goldinger, 2000). Twelve subjects performed the baseline and training phases, then were re-recorded saying all words one week later (without shadowing). These productions were found to have shifted towards the training productions, as measured by AXB similarity. Individual results are not reported in either study (1998, 2000).

Following Goldinger and other work on spontaneous imitation in laboratory settings (Fowler et al., 2003; Namy et al., 2002), an influential study by Pardo (2006) examined short-term shifts during social interaction. Six same-gender dyads participated in a map task,[29] and were recorded saying each landmark's label in pre and post-task sessions. A panel performed AXB tasks resulting in two types of similarity score: a measure of shift

[28] The actual experiments involve a range of manipulations which are not relevant here.

[29] In this task, two speakers separated by a divider each have a map labeled with landmarks. One speaker (the giver) has a path on his map while the other (the receiver) does not. The task is for the receiver to duplicate the path on his map using directions from the giver. The resulting interaction leads to many repetitions of the landmark names by each speaker (Anderson et al., 1991).

towards the other speaker within the conversation (for within-task items), and the persistence of shifts into the post-task session (for post-task items). As measured by AXB similarity, most subjects shifted towards each other during the task, and the shifts persisted to some extent post-task. All subjects showed some shift during the task (Fig. 2), though some shifted away from their partner. Pardo reports a complex interaction between a speaker's role and sex on the size and direction of shift, and concludes that social factors affect the degree of accommodation. However, the robustness of this interpretation is unclear given the small number of speakers (12) relative to the number of by-speaker variables (3).

Pardo (2009) analyzed data for six same-sex dyads from the preceding experiment, as well as four mixed-sex pairs, for shifts in pitch and utterance duration. Shifts for both variables were moderately predictive of AXB similarity scores, suggesting some convergence in both. There was also modest evidence for shifts in speakers' vowel spaces towards each other, comparing formant values for pre-task and post-task vowels. For all three variables (f0, utterance duration, vowel formants) effects of gender and talker role are reported, but they are again hard to assess given the number of speakers considered. Pardo et al. (2010) performed a similar experiment to Pardo (2006), but with 12 same-gender dyads, and the instruction given to each speaker varied as well (one member of each dyad was explicitly told to imitate). Convergence between speakers in a dyad was again found, mediated by talker role, gender, and instruction. The largest effect was of talker role (givers shifted towards receivers). Acoustic analyses were also performed of speaking rate and vowel formants, but little evidence was found for convergence. Speakers substantially varied their speaking rates during interaction, but not towards each other.

The literature on short-term shifts in pronunciation (but where phonetic and phonological variables are not examined) is growing rapidly, and we omit discussion of several recent studies that are less relevant for our purposes, which examine pairs of nonnative speakers, pairs of native and nonnative speakers, and the role of phonetic talent on the degree of shift (Kim, 2011; Lewandowski, 2012; Lewandowski and Dogil, 2010).

4.3.2.3 Shifts in phonetic and phonological variables

We now turn to short-term studies which examine phonetic/phonological variables. Our review is limited in two ways. We do not discuss the broader sociolinguistic literature on *style shifting*, where an individual's use of a variable changes depending on the context (interlocutor, topic, formality, attention, etc.) (Schilling-Estes, 2003). We are interested only in cases where all factors except the interlocutor's identity are as controlled as possible, to understand how much short-term shift occurs due to interaction alone. In addition, we note that the literature on short-term shifts in phonetic and phonological variables is currently a very active research area, and a number of recent studies which are less relevant for our purposes are not reviewed here (Aubanel and Nguyen, 2010; Kappes et al., 2009; Kim et al., 2011; Lawson et al., 2011; Lelong and Bailly, 2011, 2012; Michelas and Nguyen, 2011; Nilsenová et al., 2009; Nilsenová and van Amelsvoort, 2010; Pardo et al., 2012; van Dommelen et al., 2011).

Voice onset time Shockley et al. (2004) examined whether subphonemic detail (VOT) would be imitated in a shadowing task. All shadowed words began with a voiceless stop, and baseline productions of each word were produced before the task by each of eight subjects. In one experiment the model talker's word-initial VOTs were doubled; in a second experiment they were not. Speakers produced longer VOTs for shadowed productions relative to baseline productions in both experiments, but the difference was larger in the lengthened-VOT experiment, indicating that speakers imitated lengthened VOT. No individual results are reported.

Nielsen (2011) examined VOT imitation using a modified version of Goldinger's paradigm. Subjects produced a list of words beginning with /p/ or /k/, then heard a subset of the /p/ words produced by a model speaker with the VOTs artificially modified; they then produced all words again. The speaker's VOTs were lengthened in one experiment, and shortened in a second experiment. In the first experiment, subjects' VOTs for post-listening productions were lengthened relative to baseline productions. Of the 27 subjects, 25 increased their average VOT, some much more than others (Nielsen's Fig. 2). In contrast, in the second experiment, there was no significant change in between the baseline and post-listening productions. VOT increased for 14 subjects and decreased for 11 subjects. Nielsen concludes that subjects will imitate lengthened but not shortened VOT in voiceless stops, possibly because decreasing VOT would endanger the contrast with voiced stops.

Abrego-Collier et al. (2011) examined VOT imitation for word-initial voiceless stops in a task similar to Nielsen's in overall structure (baseline, listening, post-listening), but with a listening phase where 72 target words with lengthened VOTs were embedded in a narrative about a blind date, rather than presented in isolation. Several social variables (e.g., subject gender, talker sexuality), as well as subject attitude towards the narrator were examined for effects on the amount of change in VOT for 48 subjects. Only attitude had a significant effect: VOT increased for subjects who felt positively towards the narrator (towards his lengthened VOTs), and decreased for subjects who felt negatively. In contrast to Nielsen's finding, there was no overall shift in VOT across subjects. Individual differences are not reported in the paper, but were later examined for an expanded dataset of 87 speakers. Two models were fit for each speaker's data, one with terms indexing baseline vs. post-listening blocks, and one without them. Comparing the two models gives a significance value, which measures how confident we are that speakers showed any change in VOT between blocks. VOT changed significantly ($p<0.1$) for 56 subjects (64%).

An additional study discussed below investigates imitation of negative VOT (Mitterer and Ernestus, 2008).

Vowels Imitations are of course never exact: for example, when imitating vowels in a laboratory setting, speakers' imitations are reliably noisy (variable across imitations) and biased (systematic difference between imitations and the target) (e.g., Chistovich et al., 1966; Repp and Williams, 1985). Vallabha and Tuller (2004) carefully studied the nature of inaccuracies in imitation of self-produced vowels spread throughout the F1/F2 space, by three speakers. Imitations were highly biased and noisy, but the exact pattern differed greatly across speakers. However, speakers had similar distributions for the *amount* of bias and noise, suggesting that they imitated with equal degrees of inaccuracy. Speakers' imitation patterns suggest "there is additive Gaussian noise in the imitation process that independently affects each formant" (1184). Finally, of two speakers who repeated the experiment 14 months later, one showed significant change in imitation patterns and one did not, suggesting that whatever factors underlie an individual's imitation patterns can change over time.

Delvaux and Soquet (2007) examined phonetic convergence by speakers from one French dialect region of Belgium to a recorded speaker from a different dialect region. Subjects performed a picture naming task, with no explicit instructions to imitate. They named pictures alternately with the recorded speaker, thus exposing them to her speech. Of primary interest was how speakers shifted their realization of two vowels whose phonetic realization differs between Belgian regiolects. Shifts in three types of acoustic parameter—duration, formant frequencies, and MFCCs—were assessed using an innovative methodology: a discriminant analysis was performed to determine the subspace of formant frequency space (for example) best differentiating speakers from the two dialect regions (the subject's and the recorded speaker's); convergence was assessed by whether within-task productions were less separable from the recorded speaker's productions, in this space, relative to baseline productions. In one experiment four speakers converged to the recorded speaker, in all three types of acoustic parameter. In a second experiment, only MFCCs and duration were analyzed. Robust convergence was observed for all subjects: across both variables, two vowels, and eight subjects, a shift towards the model speaker occurred in 30 of 32 cases, with the two exceptions from different speakers.

Analyses of vocalic variables in Pardo (2009; et al. 2010) are discussed above.

Babel (2010) examined shifts in vowel formants in a word shadowing task, and their conditioning by social variables. Thirty-four New Zealanders produced words containing six vowels; they then shadowed productions of these words by Australian talkers, then produced the words again. For each shadowed production of a word, the difference was calculated between (1) the euclidean distance between the shadowed production's vowel and the target production's vowel, in F1/F2 space, and (2) the distance between the baseline and target productions, by the same metric; a negative difference indicated convergence towards the target pronunciation. A similar score was calculated for the post-task productions. Overall, all vowels shifted towards the target pronunciations, some much more than others. In addition, the amount of convergence towards the Australian talker was mediated by the subject's attitude towards Australia (positive⇒more convergence). Individual difference results are not explicitly reported. However, 25 of 34 subjects seem to show overall convergence (negative average difference-in-distance scores in Babel's Fig. 1).

The same overall method (single-word shadowing, baseline/task/post-task) was used by Babel (2011) to study shifts in vowel formants by 111 American English speakers towards a male model talker, in one of four conditions: the talker was African American or Caucasian, and his face was visible or not. Words contained one of five vowels (/æ/, /ɑ/, /i:/, /o:/, /u:/), and convergence was assessed using the same distance-in-distance metric. Subjects converged more for low vowels (/æ/, /ɑ/) than for non-low vowels, and the amount of convergence varied by condition, the subject's sex, and perceived model talker attractiveness. Individual differences are not explicitly reported, but most speakers seem to have shown some shift for productions of each vowel, on average, especially the low vowels (Babel's Figs. 4, 6, 7).

Nielsen (2010) examined imitation of Japanese vowel devoicing. In Tokyo Japanese, short high vowels are variably devoiced between two voiceless consonants; there is disagreement over whether this is a categorical phonological (assimilation of [-voice]) or gradient phonetic process. The experimental setup was similar to Nielsen's VOT experiment, with the listening phase consisting of words artificially modified such that *all* vowels which could be devoiced were. Gradient and categorical measures were used to determine how much/whether subjects' productions were devoiced. Overall, subjects shifted towards the model talker, devoicing more in post-task productions than in baseline productions. Most individual speakers (20 of 24) showed this pattern, a few much more than others (Nielsen's Fig. 1).

Segmental vs. phonetic variation All studies discussed so far examined imitation of gradient, subphonemic variation. Three other studies examined imitation of segmental variation (the use of one categorical variant instead of another), as well as how much contrastive variation is imitated relative to non-contrastive variation.

Mitterer and Ernestus (2008) examined imitation of categorical and gradient variation by Dutch speakers using a word shadowing task. The categorical pattern was variation in the phonetic realization of /r/ in Dutch. The most common realizations are alveolar [r] and uvular [ʀ], and most speakers use one or the other exclusively. The gradient pattern was VOT, which in Dutch distinguishes voiced (negative VOT) from voiceless (positive VOT) stops. Eighteen subjects participated, both habitual [r] and habitual [ʀ] users. Subjects shadowed non-words beginning with /r/ or a voiced stop, where /r/ was realized as [r] or [ʀ], and the voiced stops with 0, 6, or 12 cycles of prevoicing. There was very little imitation of /r/ realization: in 97.3% of cases where the subject's habitual variant (e.g., alveolar) did not match the target realization (e.g., uvular), the habitual variant was used in the shadowed production. The 2.7% of cases of imitation were concentrated in two subjects; all others almost never imitated. Imitation for voiced stop stimuli was assessed using the duration of prevoicing in the shadowed production. Subjects produced longer prevoicing when shadowing stimuli with prevoicing (6 or 12 cycles), compared to those without prevoicing, but did not differ in the duration of prevoicing when shadowing stimuli with 6 or 12 cycles of prevoicing. (Individual results are not reported for the stop stimuli.) Thus, "the phonologically relevant difference between presence and absence of prevoicing was imitated, while the phonologically irrelevant amount of prevoicing was not" (173). Assuming that speakers treat different realizations of /r/ as equivalent, its realization is also "phonologically-irrelevant." That speakers did not imitate non-contrastive aspects of phonetic realization supports Mitterer and Ernestus's view that "the link between speech perception and production is phonological and abstract" (168).

Brouwer et al. (2010) investigated how listeners imitate reduced speech. Sixteen subjects shadowed canonical and reduced pronunciations of words excised from a corpus of spontaneous spoken Dutch; six subjects were excluded based on poor performance. Imitation was assessed using the duration of the shadowed response and how its constituent segments were realized (the same as or differently from the target's realization). Shadowed productions were longer than target productions overall, with a larger difference for reduced targets. Subjects rarely shadowed canonical segments as reduced; that is, the shadowed production was rarely more reduced than the target. However, they often (68%) shadowed reduced segments as canonical. Thus, canonical forms were imitated more closely than reduced forms, in terms of both duration and segment realization, perhaps because "listeners reconstruct canonical forms from their reduced forms" (36). No individual results are reported.

Cole and Shattuck-Hufnagel (2011) examined whether listeners imitated prosodic categories (accents and boundaries), and the phonetic realization of particular categories. Six subjects repeated utterances excised from a spontaneous speech corpus of American English. Both the target and repeated productions were prosodically transcribed for the presence of accents and boundaries, in ToBI format. Phonological form was usually preserved: accents or boundaries were seldom deleted or added, though all speakers did so to some extent. In contrast, a preliminary analysis of some data suggested that the phonetic realization of prosodic categories was not imitated. Thus, subjects imitated contrastive but not non-contrastive variation, as for Mitterer and Ernestus (2008).

Sociolinguistic variables All studies discussed so far address short-term shifts in experimental settings. A final strand of studies analyzes short-term shifts during conversational interaction in traditional sociolinguistic variables: categorical variables which are socially stratified (e.g., by region, social class, race).

Coupland (1984) examined how a shop assistant in Cardiff (Wales), "Sue", uses four categorical sociolinguistic variables in interactions with 51 customers from different social classes. Each variable is socially stratified: higher-class customers use fewer non-standard variants. With customers pooled by class, Sue clearly accommodated towards customers in each variable: correlations between her rate of usage of the non-standard form and the customers' range from 0.76 to 0.90.

Trudgill (1986: 7–11) examined the rates of use of two sociolinguistic variables (the realization of BATH and of /t/) by himself and the interviewee in 10 sociolinguistic interviews conducted a decade earlier in his hometown (Norwich, England), to test whether accommodation took place. There is strong evidence for accommodation for /t/, due to Trudgill accommodating to the interviewees usage, but not for BATH. The different behavior of the variables is ascribed to the high social salience of /t/ relative to BATH.

Hay et al. (1999) analysed the use of the socially-salient /ay/ vowel by the talk show host Oprah Winfrey for effects of her interlocutor's ethnicity and lexical frequency. Oprah used the monophthongal [a:] variant (which is associated with African American English) significantly more frequently with African American guests than with non-African American guests, suggesting she was accommodating based on the guest's ethnicity.

Two other short-term studies examine short-term shifts in communities where leveling of local dialects towards a (regional or national) standard is taking place. The question in each case is whether community-level leveling is foreshadowed in short-term shifts during interaction, as predicted by an influential theory of the sources of dialect change (see Sec. 4.3.3.1)

As part of a larger study of dialect leveling in Luxembourgish, Gilles (1999, summarized in Auer and Hinskens, 2005) examined the use of four vocalic features by five speakers of non-standard dialects of Luxembourgish (only some variables for each speaker), in conversations with speakers of other dialects (who use the standard variants almost exclusively) vs. speakers of their own dialect. In most cases speakers used *fewer* standard forms in interdialectal conversations than in intradialectal conversations, the opposite of the predicted pattern. The amount of shift (difference between interdialectal and intradialectal usage) varied hugely between speakers and variables, with the opposite (i.e., predicted) pattern sometimes occurring.[30] The results for one variable suggest 'psychological accommodation' towards a stereotype of the variable's usage in the conversation partner's regional dialect, rather than her actual usage during the conversation (Thakerar et al., 1982).

As part of a larger study of dialect leveling in Limburg (a Dutch-speaking region), Hinskens (1996) examined the use of three consonantal features (ɣ$^{\dashv}$-weakening, n-deletion, t-deletion) by 27 speakers of the dialect of the village of Rimburg, in conversations both with other Rimburg speakers (C1; all 27 speakers) and with speakers of three progressively more standard varieties of Dutch (C2, C3, C4; nine speakers each). The non-standard variants of the three features have different areal spreads: ɣ$^{\dashv}$-weakening only in C1, n-deletion in C1/C2, and t-deletion in C1/C2/C3. For each variable, the standard variant was used less in in-group conversations (C1) than in out-group conversations (C2,

[30] From the empirical data it looks like some shift nearly always occurs, but statistical significances are not given.

C3, or C4), the expected pattern. (The difference was significant for two variables and marginal for one.) The difference was greatest for the most local variable (ɣ˧-weakening), suggesting that more geographically-restricted variants were more susceptible to short-term shifts.

4.3.2.4 Discussion

The short-term literature sheds some light on the three questions raised above about shifts in pronunciation, on the timescale of laboratory experiments, conversations, or sociolinguistic interviews.

Stability and change Short-term shifts were found for both gradient variables (VOT; vowel formants, duration, overall spectrum, and amount of devoicing; segment reduction) and categorical variables (prosodic realization, segment deletion, several sociolinguistic variables). A smaller number of variables showed marked stability, most strikingly Dutch /r/ realization. For variables which did change, when individual differences were reported, there were large differences in the magnitude of shift for different speakers, who sometimes differed in the direction of the shift as well. But crucially, *most speakers showed some shift*. We do not observe a majority of speakers who are stable and a minority who change, as in many long-term studies. In some short-term studies (e.g., Babel, 2011; Nielsen, 2011) a small number of speakers changed much more than others, but overall we do not see a division between changers and non-changers.

Interspeaker differences In most studies where individual results were reported, the amount of short-term shift varied significantly among speakers. In some cases social factors provided a simple and intuitive explanation for some variation, for example when the direction and degree of shift lined up well with feelings towards the interlocutor (Abrego-Collier et al., 2011; Babel, 2010), or the interlocutor's dialect region (Delvaux and Soquet, 2007), social class (Coupland, 1984), or use of the variable (Trudgill, 1986).

However, in other studies the proposed role of social factors is more complicated. For example, Pardo (2006) found that women shift less than men, direction givers shift towards direction receivers, and the two factors interact; Pardo proposed possible explanations related to speakers' communicative goals. However, the opposite pattern was expected from the Communication Accommodation Theory literature: women generally shift towards men, and receivers towards givers (e.g., Namy et al., 2002). Other short-term studies have found conflicting gender effects: Abrego-Collier et al. (2011) found no effect on VOT, and Babel (2011) found a complex interaction with other factors (but no main effect) on vowel formants. In each case, a reasonable account is given of why social variables affected short-term shifts in the observed manner, and why some predictions from previous work were not met. But the existence of such different findings suggests that the effects of social variables on shifts in pronunciation are far from straightforward, raising the question of how confident we can *ever* be in the post-hoc account given for a particular case.

One response to this issue notes that the link between social variables and shifts in pronunciation is expected to be complicated. Even in a controlled setting, how speakers shift will depend on the experimental setup (e.g., conversation, a map task, shadowing), social dynamics with the interlocutor, the topic of conversation, and so on; whole subfields of sociolinguistics (style shifting, discourse analysis) examine such shifts during conversation. Much more study is needed before we can make valid generalizations about the factors affecting short-term pronunciation shifts in experiments. The researcher's job is to perform a detailed analysis of an experiment's results, and find a plausible interpretation. Theories from social psychology or anthropology are invaluable in doing so, when used responsibly.

Another response questions the utility of such theories in general in explaining patterns of shift. Meyerhoff (1998: 208) laments that CAT, the most frequently-used framework, is often invoked in the face of messy data "to give the impression that the investigator has 'explained' all observed patterns." Siegel (2010: 73) notes that "many of [CAT's] claims do not seem to be falsifiable... the theory predicts that in an interaction, a person with less power will converge with (or accommodate to) a person with more power. But the theory also stipulates that this prediction can be overridden by other situational or personal factors."[31] These criticisms apply equally to automatic accounts: imitation is automatic if observed, and mediated by a social variable if not.

In our view, these critiques are very important, and it is still not clear what can be concluded about the effect of social variables on short-term shifts in phonetic and phonological variables. Importantly, for our purposes what matters is not the interpretation of these shifts, but their *existence*, and how general they are across speakers and variables.

Intervariable differences Many short-term studies have found differences in how susceptible different variables are to shift, and propose a range of explanations.

Phonological contrast: Mitterer and Ernestus (2008) suggests that short-term shift is phonologically mediated: speakers will imitate contrastive but not subphonemic variation. This view is also supposed by Cole and Shattuck-Hufnagel's pilot study of prosodic imitation. However, many other studies (e.g., Babel, Nielsen, Delvaux & Soquet) have found robust short-term shifts in subphonemic variation. Nielsen (2011) suggests that contrast maintenance is a factor in what short-term shifts are possible: speakers shifted towards increased but not decreased VOT, because decreasing VOT would endanger the contrast with voiced stops. However, speakers in many other studies, including Nielsen's (2010) work on vowel devoicing, shift in contrast-endangering directions.

Social salience: Trudgill (1986) suggests that socially-salient variables are more likely to undergo both short-term and long-term shifts. However, this pattern has not been

[31] Note that Siegel and Meyerhoff are discussing the use of CAT in studies of second dialect acquisition and sociolinguistics, respectively. However, their criticisms carry over directly to experimental studies of short-term shifts in pronunciation.

found in other studies where several variables of different degrees of salience are considered. Trudgill also predicts that variables which are *too* salient ("extra-strong salience") will not shift, a move widely criticized as making the definition of salience circular (e.g., Hinskens, 1996; Kerswill and Williams, 2002). Babel (2010) found that the most socially-salient vowels shifted the least for her New Zealand English talkers, and suggests their salience *prevented* them from shifting. Vallabha and Tuller (2004) found that speakers shifted in response to target vowels throughout formant space (hence, regardless of social salience), though the role of social salience was not explicitly examined.

Distance from target: Babel (2011) found that speakers converged much more for low vowels than for non-low vowels. Since the realization of low vowels varies greatly among American English dialects, she speculates that speakers may have reference vowels further from the target on average, and thus have more room to shift (however, information on speakers' dialect regions was not available). Similarly, Babel (2010) found that two of the vowels which showed the most convergence were those which differ the most between Australian (model talkers) and New Zealand (subjects) English.

Geographical spread: Considering three categorical variables, Hinskens (1996) found that the amount of shift in a variable by speakers of a regional Dutch dialect, in conversation with speakers of other dialects, was inversely related to the variable's geographical spread. This can also be interpreted as the amount of shift in a variable correlating with the difference in its rate of usage between the two speakers, which might be interpreted as "distance from target" for categorical variables.

In sum, many explanations have been proposed to explain differences in the degree of short-term shift for different variables, but a coherent picture has not yet emerged. Most explanations have not been tested in more than one or two studies, or have several exceptions. Unlike in the long-term literature, there have been no short-term studies considering a large number of variables of different types at once, so the gamut of proposed explanations have never been compared on a fixed dataset. As more studies are carried out, we speculate that a variable's short-term plasticity will depend on a range of factors, as is the case for long-term change.

4.3.3 Linking short-term and long-term

4.3.3.1 The change-by-accommodation model

Beginning with the Neogrammarians (Paul, 1880), variants of what Auer and Hinskens (2005) call the *change-by-accommodation model* (CBA) have been proposed to explain the relationship between individual-level and community-level change, as a sequence of three steps:[32]

1. *Short-term change*: People accommodate to each other during interactions.

2. *Long-term change*: Eventually, an individual's norms change as a result of these interactions.

3. *Language change*: Spread of the innovation in the wider community.

We focus on Steps 1 and 2, which describe change in the individual. The best-known statement of CBA comes from Trudgill's seminal work on dialect change, and invokes accommodation theory:

> In face-to-face interaction... speakers accommodate to each other linguistically by reducing the dissimilarities between their speech patterns and adopting features from each other's speech. If a speaker accommodates frequently enough to a particular accent or dialect, I would go on to argue, then the accommodation may become permanent, particularly if attitudinal factors are favourable... (Trudgill, 1986: 39)

[32] Auer and Hinskens use the terms "short-term accommodation" and "long-term accommodation."

In Trudgill's terms, short-term accommodation, if it occurs often enough, leads to long-term accommodation. In other words, the more you interact with speakers who speak differently from you, the more your speech shifts towards theirs over time. Variants of CBA have been proposed both for dialect change/contact and for change internal to a speech community. Two basic versions of the model can be distinguished, corresponding to the automatic and social positions on the causes of short-term shifts.

One view is that the short-term shifts which happen during conversation are largely subconscious and automatic, and identity and social factors play a much smaller role. Higher use of some variant in a social group comes about as a result of more frequent contact among members of the group than with non-members, not because the variant serves a social function. Trudgill (2004) has most forcefully argued this viewpoint, in the context of new dialect formation. Much earlier, Bloomfield (1933) emphasized the importance of the *principle of density*: every time two speakers communicate, their sound system becomes more similar.[33] Citing the principle of density and speaking more guardedly than Trudgill, Labov (2000: 506) argues that "it is good practice to consider first the simple and more mechanical view that social structure affects linguistic output through changes in the frequency of interaction" rather than assume social motivations for Steps 1–2 by default. Finally, the automatic view is explicitly adopted in some short-term studies: for Delvaux and Soquet (2007: 146), the motivation for studying the effect of exposure to speech from a different dialect region is "the hypothesis that sound change partly originates in mutual and unintentional imitation between interacting speakers."

On another view, short-term shifts during conversation are fundamentally social, as proposed by CAT, the result of accommodation by speakers to achieve social goals. Under this view, the social stratification of speech results from variants which have become

[33] However, Bloomfield also emphasized the role of (what is now called) prestige in which short-term shifts occur: "The humble person is not imitated; the lor or leader is a model to most of those who hear him. In conversation with him, the common man avoids giving offense or cause for ridicule; he suppresses such of his habits as might seem peculiar, and tries to ingratiate himself by talking as he hears" (476–477).

associated with a particular social groups being used more by those who identify with the group and less by others, and so on. The role of social factors has already been alluded to in Trudgill's "particularly if attitudinal factors are favourable" above. In the short-term literature, results which show effects of social variables on short-term shifts are sometimes interpreted in the context of a social version of CBA. For example, Pardo (2006) interprets the finding of gender and talker role effects on shifts in pronunciation as "rapid phonetic convergence that emerges in conversational settings, providing a link between the laboratory studies of nonsocial shadowing imitation and community-level linguistic change."

The identity projection model Auer and Hinskens (2005) call into question a key assumption of CBA: that people shift with respect to the interlocutor's speech (i.e., directly towards or away from it) in Step 1. They discuss a number of short-term studies, including the Luxembourg study discussed above, which suggest that speakers are shifting not with respect to their interlocutor, but a subjective picture of how he 'should' sound. More generally, the *identity-projection model* holds that short-term shifts in conversation result from the speaker expressing a particular persona by shifting his usage towards what he thinks that persona sounds like. Thus, frequency of interaction or face-to-face contact with interlocutors who use a variant are not necessary for a speaker to increase his use of the variant over time.[34] Auer and Hinskens point out that the subjective component of accommodation has long been recognized in the CAT literature. In response to cases where divergence was observed where convergence was expected, a distinction between 'objective convergence' and 'subjective convergence' was introduced into CAT (Thakerar et al., 1982).

The identity projection model assumes that long-term change does occur in individuals (Step 2), and foreshadows change in the community (p. 351). Long-term change is not

[34] The identity projection model has much in common with Le Page and Tabouret-Keller's 'acts of identity' and Bell's 'referee design' (Bell, 2001; Le Page and Tabouret-Keller, 1985).

explicitly stated to result from an accumulation of short-term shifts (towards a subjective target), but this seems to be assumed in the discussion.[35]

4.3.3.2 Medium-term change

Our review of the short-term and long-term literatures suggests two motivations for studying *medium-term* change: intermediate timescales between hours and years.

The mismatch between short-term and long-term change Both CBA and the identity projection model make an important assumption, which we call the *persistence hypothesis*: that short-term shifts which an individual makes during interaction can and do accumulate over time into long-term change. The short-term literature discussed above suggests that short-term shifts do take place, for most variables studied so far, for most adult speakers. Thus, if there is any systematicity in a speaker's short-term shifts for a variable during interaction—whether to the interlocutor's speech, or to a stereotype—his use of the variable should change over the long term. But this conflicts with the long-term literature, where we saw that most speakers are stable, while a minority shift significantly. There were also huge differences in how plastic different variables were in the long term, while nearly all variables in short-term studies show shifts. How can we reconcile the different patterns seen in short term and long-term change, and where does the disconnect between the two lie? To understand the mismatch between short-term and long-term dynamics, we must understand what longitudinal variation in individuals looks like over timescales in between.

Problems in interpreting short-term and long-term studies Another motivation is that we know very little about any timescale between short-term and long-term studies—yet what happens over intermediate timescales is crucial for interpreting the results of both.

[35] cf. Hinskens (1996: 20): "Like Bloomfield and Trudgill, we are convinced that the process of dialect leveling starts as the interactionally motivated phenomenon of linguistic accommodation."

The persistence hypothesis is tacitly assumed by much short-term work, where a major motivation is to understand the first steps of language change under a CBA-style account. However, to my knowledge there is no work examining the persistence of short-term shifts in phonetic or phonological variables for more than one hour.[36] Thus, we actually have no hard data on whether such shifts can persist over days to months. If the persistence hypothesis is false, short-term shifts would be interesting phenomena, but with limited relevance to language change.

In long-term studies, an individual is recorded at several timepoints (usually two) years apart to see whether change occurred in some variables. This procedure assumes that individuals vary relatively little from day to day in their use of a variable. Under this *stationarity hypothesis*, any change observed in a speaker's use of the variable after controlling for conditioning factors is meaningful. But as Nahkola and Saanilahti pointed out, we know very little about whether individuals fluctuate from day to day in their use of most variables.[37] If the stationarity hypothesis is false, understanding how much speakers fluctuate from day to day is crucial for interpreting the results of long-term studies.

Finally, at least in their simplest forms, the stationarity hypothesis and the persistence hypothesis conflict. If short-term shifts are persistent, than there would be significant noise even within the same day, violating the stationarity hypothesis. If the stationarity hypothesis holds, there can be no accumulation of short-term shifts. Which hypothesis, if either, is correct? Both can be checked by examining trajectories of variables within individuals over the medium term. If they show clear time trends, then these *could* be the result of accumulated short-term shifts. If they show little day-to-day noise, the stationarity hypothesis is supported.

[36] Goldinger (2000) shows persistence over 1 week, but of the overall percept rather than an acoustic dimension.

[37] One exception is a laboratory study of vowel productions by two American English speakers on two consecutive days Pisoni (1980).

4.4 Previous work on phonetic variables

4.4.1 Voice onset time

VOT was introduced In Sec. 3.1. A vast literature on VOT in speech production and perception has developed over the past 50 years, from both experimental and clinical perspectives. Here we review important aspects of previous speech production studies of VOT in the (experimental) phonetics literature, which is the strand most relevant to the current study.[38] Most production studies have examined VOT in planned speech in laboratory experiments, in contrast to our study of VOT in spontaneous non-laboratory speech.[39] In particular, VOT is most often examined for stops in words or syllables spoken in a laboratory setting, either in isolation or in a carrier phrase. Fewer studies have examined VOT in read speech (Byrd, 1993; Crystal and House, 1988a,b; Lisker and Abramson, 1967; Picheny et al., 1986; Randolph, 1989) or in conversational speech (Baran et al., 1977; Lisker and Abramson, 1967; Yao, 2009b). Two studies have built models to understand the joint effect of different factors on VOT in speech corpora: Yao applies linear regression to VOTs from the Buckeye Corpus of spontaneous speech (2 speakers, 776 VOTs), and Randolph builds a regression tree for 7900 VOTs from three read speech databases (including TIMIT).[40] Both studies reach interesting conclusions relative to laboratory studies. Yao finds that her two speakers differ greatly in which factors affect VOT the most, and that the most-discussed factor in laboratory studies (place of articulation) has a small effect relative to other factors. Both studies find that much of the variation in

[38] Even this strand is too large to give a full review here, and to my knowledge none exist. Useful partial reviews are given by Auzou et al. (2000) and Yao (2009b), and Docherty (1992) gives a comprehensive review of studies of VOT in English up to about 1990.

[39] A further difference is that most studies have considered American English speakers, though there is a respectable sub-literature on VOT in British dialects, mostly Received Pronunciation (e.g., Docherty, 1992; Docherty et al., 2011; Heselwood and McChrystal, 2000; Masuya, 1997; Scobbie, 2006; Suomi, 1980; Whiteside and Irving, 1997). Because the Big Brother speakers come from a variety of dialect regions and are not all British, we hesitate to call this a study of VOT in British English. The main difference from previous work is that almost all speakers are not American.

[40] Randolph actually considers stop release duration, a slightly different measure from VOT.

Table 4.1: Expected effects of conditioning factors on VOT for English word-initial voiceless stops.

Factor	Expected
Place of articulation	/p/ ≤ /t/ ≤ /k/
Following segment type	Vowel < Sonorant
Following vowel	Non-high < High
Syllable stress	Unstressed < Stressed
Pitch	Low/medium < High
Speaking rate	Faster < slower
Gender	Male ≤ female
Individual speakers	Differ in baseline VOT Differ in effect of speaking rate Do not differ in effect of POA

their data remains unexplained by the model.

4.4.1.1 Conditioning factors

There is huge VOT variation in stops produced by English speakers, much of which is due to properties of the linguistic context, speaker, type of speech, and the recording environment. We review the major conditioning factors, specifically in word-initial voiceless stops (which our dataset consists of), with more detail given for factors which will be modeled for the Big Brother data. Table 4.1 summarizes the expected effect of each factor. We will not discuss proposals as to *why* each factor influences VOT in a particular way. However, many proposals exist, and can be found in the references cited.

Segmental properties Whether a stop consonant is voiced or voiceless is probably the largest single factor affecting VOT.[41]

After voicing, the best-known factor influencing a stop's VOT is its place of artic-

[41] This intuition is supported by the results of Randolph's model, where voicing specification is the first split in the decision tree, implying that this feature was more informative for determining VOT than any other single feature. However, his model does not include speaking rate, which would be expected to have a very large effect.

ulation: VOT increases for stops articulated more towards the back of the vocal tract, cross linguistically (Cho and Ladefoged, 1999; Lisker and Abramson, 1964). The expected ordering for voiceless stops in English (and cross linguistically) is /p/</t/</k/ (e.g., Lisker and Abramson, 1967; Nearey and Rochet, 1994; Zue, 1976). Other studies have found a subset of the total order: /p/=/t/</k/ or /p/</t/=/k/ (e.g., Cooper, 1991; Crystal and House, 1988a; Docherty, 1992).

VOT is also influenced by the identity of segments following the stop consonant. Initial stops in consonant clusters (e.g., /pr/, /kl/) tend to have longer VOTs than those in CV syllables (Docherty, 1992; Klatt, 1975; Morris et al., 2008; Zue, 1976). The quality of the following vowel (whether the stop occurs in a consonant cluster of not) has also generally been found to have some effect (except by Lisker and Abramson, 1967; Zue, 1976) A generalization describing most previous work is that VOT for English voiceless stops is lengthened before high vowels, especially /i/ and /u/, compared to non-high vowels (e.g., Docherty, 1992; Klatt, 1975; Nearey and Rochet, 1994; Ohala, 1981; Port and Rotunno, 1979), though different studies have found a variety of (sometimes conflicting) more specific effects. Some studies have found an interaction between the effect of the following vowel and the stop's place of articulation on VOT (e.g., Morris et al., 2008; Summerfield, 1975b), while others have not (e.g., Nearey and Rochet, 1994).[42]

Suprasegmental factors VOT decreases with increased speaking rate for voiceless stops, but shows little or no change for voiced stops (e.g., Allen et al., 2003; Miller et al., 1986; Summerfield, 1975a; Volaitis and Miller, 1992). The asymmetry between the rate effect for voiced and voiceless stops appears to hold cross linguistically (in Thai, French, and English; Kessinger and Blumstein, 1997). The effect of speaking rate is large relative to other factors.[43] Accordingly, speaking rate is often controlled for in experimental studies

[42] Docherty (1992:143) reports mean VOT values, cross tabulated by place of articulation and height of the following vowel, which seem to show an interaction. However, a statistical test is not given.

[43] For example, in a study of isolated English words beginning with voiceless stops, Allen et al. (2003: 549) found that "82% of the total variability [in VOT] was attributable to differences in talkers in overall

of VOT (when it is not itself the object of study), usually using a local measure of duration (of the following vowel, or the syllable or word containing the stop) as a proxy.

Stress and pitch also affect VOT. For voiceless stops, VOT is longer before stressed vowels than before unstressed vowels, and the opposite effect holds for voiced stops (Lisker and Abramson, 1967), making VOT is a less reliable cue to stop voicing in unstressed syllables. In an experiment where speakers read sentences at one of three pitch levels, McCrea and Morris (2005) found that VOT of voiceless stops increased for high pitch, compared to medium or low pitch, but there was no pitch effect for voiced stops.

Other linguistic factors Other aspects of the host word affect VOT. In line with the general finding that segments tend to undergo more reduction in frequent words and in more predictable contexts (see Bell et al., 2009 for a recent review), one might expect that VOT would be shorter for more frequent words. VanDam and Port (2005) and Yao (2009b) found significant effects in the expected direction, while Abrego-Collier et al. (2011) found no significant effect.[44] VOT is longer for word-initial stops in monosyllabic words, relative to disyllables (Klatt, 1973; Lisker and Abramson, 1967). For closed syllables beginning with a stop, VOT is longer when the coda is voiced vs. voiceless (Port and Rotunno, 1979; Weismer, 1979).

Finally, VOT is affected by properties of the utterance the host word appears in. Lisker and Abramson (1967) and Baran et al. (1977) found that voiceless stops in isolated words generally have longer VOTs (i.e., more positive) compared to words in sentences, and the opposite effect holds for voiced stops; that is, the voicing distinction is enhanced for words spoken in isolation.[45] Docherty (1992:140) observes a similar pattern, but it does

rate". It is not totally clear how this figure is obtained, but Allen et al. 's point, that rate is by far the most important explanatory factor for their data, holds.

[44] In fact, Abrego-Collier et al. found a marginal effect ($p < 0.1$) in the opposite direction: more frequent words had *longer* VOTs.

[45] Lisker and Abramson considered isolated words, "minimal pairs", and words in sentences; Baran et al. considered isolated words, read speech, and both adult- and child-directed conversational speech.

not reach significance.

Speaker factors Facts about the speaker have also been shown to affect VOT. Two studies have examined the extent to which speakers differ in VOT and in the effect of factors conditioning VOT, using mixed-effects models.[46] Allen et al. (2003) found that American English speakers have different characteristic VOT values, even after controlling for speakers' different characteristic speaking rates. Theodore et al. (2009) replicated this result, and also found that speakers differed in the effect of speaking rate. (VOT decreases more with increased rate for some speakers than for other.) However, speakers did not differ in the effect of place of articulation (specifically, /p/ vs. /k/).

There have been conflicting findings on how and whether the speaker's sex affects VOT (see Morris et al., 2008). Many studies of American and British English have found that female speakers tend to have higher VOTs than male speakers (e.g., Robb et al., 2005; Swartz, 1992; Whiteside and Irving, 1997). However, as pointed out by Morris et al., these studies have generally not controlled for speaking rate.[47] Studies where speaking rate is considered (Allen et al., 2003; Morris et al., 2008), as well as some other studies (e.g., Ryalls et al., 2004, 1997; Sweeting and Baken, 1982), have not found gender differences.

Other social factors have been found to affect VOT, especially age and ethnicity (e.g., Ryalls et al., 2004, 1997; Sweeting and Baken, 1982). I omit discusion of this work, since social factors besides gender cannot be modeled for the Big Brother data due to the skewed social characteristics (e.g., speakers come from a narrow age range) of the housemates.

A final factor affecting VOT for British speakers in particular is dialect background. Heselwood and McChrystal (2000) found that English phonologically-voiced ([+voice]) stops were prevoiced (negative VOT) much more often by Panjabi-English bilingual chil-

[46] These are to my knowledge the only previous studies using mixed-effects models of VOT variation.

[47] Morris et al. do not discuss why speaking rate is a potential confound; presumably it is because women speak slightly slower than men on average (in American English, tentatively in British English: Byrd, 1994; Jacewicz et al., 2009; Whiteside, 1996), and speaking rate is negatively correlated with VOT. An apparent gender effect might therefore actually be due to the female subjects speaking slower.

dren than English monolingual children, while there was no VOT difference for voiceless stops. Scobbie (2006) finds a fascinating pattern of interspeaker VOT variation in the Shetland Islands: the percentage of prevoicing of [+voice] is inversely correlated with the duration of VOT in voiceless stops, across speakers. Thus, VOT remains the primary cue to [voice] for all speakers, but is used in different ways.[48] Docherty et al. (2011) also find interspeaker variation in VOT systems, in four towns along the Scottish-English border. [+voice] stops are realized as prevoiced or short-lag by older speakers, but almost exclusively as short-lag by younger speakers; voiceless stops have very slightly longer VOT for younger than for older speakers. Thus, there is less evidence for contrast maintenance (which would entail younger speakers having significantly higher VOT for voiceless stops) than in Scobbie's study.

4.4.2 Coronal stop deletion

Coronal stop deletion (CSD) is a variable phonological process in English in which word-final coronal stops (/t/ and /d/) are sometimes deleted in word-final consonant clusters (e.g., *bes'* vs. *best*).[49] Beginning with a series of seminal studies by Labov and colleagues (Labov and Cohen, 1967; Labov et al., 1965, 1968), CSD has been the subject of dozens of studies spanning nearly 50 years in many varieties of English. CSD has been found in every variety studied, and indeed very similar processes operate across varieties of Dutch (one of English's closest relatives) as well (e.g., Goeman, 1999; Hinskens, 1996), suggesting it has a long history. However, the vast majority of studies have considered North American varieties. For British varieties in particular, to my knowledge there have been

[48] Bradford is a city in West Yorkshire (England) where a substantial minority of the population is of South Asian origin. The Shetland Islands lie northeast of Great Britain, between Scotland and Norway. The local varieties of Scots and Scottish English traditionally show some Nordic influence, but have been rapidly changing with immigration from the mainland.

[49] That processes such as CSD are properly 'phonological' rather than 'phonetic' has been vigorously argued for over the past 15 years (see Coetzee and Pater, 2011). Nonetheless, we will often use the adjective 'phonetic', following the discussion in Sec. 1.2.

Table 4.2: Summary of expected effects of conditioning factors discussed in the text on coronal stop deletion rate. Asterisks denote a factor modeled for the Big Brother data.

Factor	Expected	Note
Following context*	Consonant > {Vowel, Pause}	Varies by variety
Preceding context*	/s/ > Sonorant > Obstruent	
Morphological class*	Monomorpheme > Semi-weak past> Weak past	No effect found by TT
Word frequency*	Higher ≥ Lower	Varies by study
Speaking rate	Faster > Slower	
Cluster voicing*	Homo-voiced > Hetero-voiced	
Gender*	Female > Male	
Situational context	Casual > Formal	
Attention to speech	Monitored > spontaneous	

two studies, on adults (ages 16–91) in York, England (Tagliamonte and Temple, 2005), and on children and caregivers in Buckie, Scotland (Smith et al., 2009). We abbreviate these studies by TT and Smith et al.

4.4.2.1 Conditioning factors

Here we summarize the major linguistic and social factors conditioning the rate of CSD. Given the diversity of accents in the Big Brother house, we pay particular attention to what is known about how consistent each factor's effects are across varieties of English. Table 4.2 summarizes the expected effect of each factor. We do not provide a comprehensive review of the CSD literature, much less the large literatures on related topics such as the determinants of segmental reduction in general. Useful recent reviews of the CSD literature are given by Schreier (2005), TT, Smith et al., and especially Hazen (2011).

Following phonological context Most studies have found the most important factor conditioning CSD rate to be the identity of the following segment, which we call the *following (phonological) context*.[50] The basic finding is that deletion occurs much more often

[50] Notable exceptions are Chicano English in Los Angeles (Santa Ana, 1996) and Tejano English in San Antonio (Bayley, 1994), where preceding phonological context is the most important factor.

before consonants than before vowels. Varieties differ in the effect of a following pause (Schreier, 2005:203–208). For example, a following pause behaves similarly to a following vowel (inhibiting deletion) in Philadelphia, but similarly to a following consonant (promoting deletion) in New York (Guy, 1980). More fine-grained categories than C/V/pause have been argued to influence deletion rate, roughly according to sonority: deletion occurs less often before less sonorous segments (glides, vowels) than before more sonorous segments (obstruents). Only the broader classification of following context (C/V/pause) is used in our models.

Preceding phonological context CSD rate is also influenced by the preceding consonant (in the word-final cluster), or *preceding (phonological) context*. Preceding context has usually been found to have a smaller effect than following context. The broad finding of previous studies on American and British varieties has been that deletion occurs more often following sonorants (glides, liquids, nasals) than following obstruents (stops, fricatives, affricates), with some exceptions. Many studies separate out a class of consonants which are somehow similar to coronal stops, and tend to favor deletion most of all, for example /s/ (most studies), all sibilants (/s/, /z/, /tʃ/, /dʒ/; TT), or /s/ and /n/ (Hazen, 2011). However, the pattern is reversed in a number of contact-induced varieties, with preceding sonorants promoting deletion relative to preceding obstruents (see Schreier, 2005). More fine-grained hierarchies have been argued for, roughly according to sonority (greater sonority ⇒ less deletion, for US and UK varieties), but are not discussed here. Our models use the broader division of preceding context used by TT (sibilants/other obstruents/sonorants).

Morphological class Along with phonological context, the most-discussed conditioning factor for CSD is the morphological class of the host word.[51] The classic finding is that

[51] Besides the empirical studies discussed here, there has been much interest among sociolinguists and phonologists in *explaining* morphological effects on CSD (e.g., Coetzee and Pater, 2011; Guy, 1991).

deletion occurs less frequently for bimorphemes (past tenses, passive adjectives) than for monomorphemes. More specifically, the expected pattern is for deletion to occur more often for regular (or "weak") past tense forms (*missed*, *walked*) than for monomorphemes (*mist*, *best*), with the deletion rate for irregular (or "semi-weak") past tense forms (*kept*, *dreamt*) falling somewhere between the two. This pattern has been found in most cases, including many regional and ethnic varieties in the US (e.g., Fasold, 1972; Guy, 1991; Hazen, 2011; Santa Ana, 1992; Wolfram, 1969), and the variety of Smith et al.'s Scottish adults.

Many contact-induced varieties of English show the opposite pattern, with word-final t/d deleting *less* in monomorphemes than in bimorphemes, for example Jamaican Mesolectal Creole, St. Helena English, Tristian da Cunha English, and Bequia English (Daleszynska, 2011; Patrick, 1991; Schreier, 2005). This pattern is likely due to the interaction of CSD with another variable process active in these varieties, where overt tense marking on past tense forms is optional. Bimorphemes are affected by both processes, while monomorphemes are only affected by CSD, so that bimorphemes show higher rates of CSD.

While the direction of the effect varies, all previous work has found some effect of morphological class, with one exception. In their study of York English, TT found no morphological effect in their statistical model, despite the presence of the expected trend (monomorphemes > semi-weak > weak) in the empirical deletion rates in their data. They suggest that the mismatch between the model and the empirical trend is due to a fact about the English lexicon: monomorphemic words are more likely than bimorphemic words to have preceding phonological contexts which promote deletion, so that "an apparently morphological effect may be an artifact of the distribution of phonological contexts across morphological categories" (296). The correlation between preceding context and morphological status was also noted by Hazen (2011), who found an unexpectedly weak (though significant) morphological effect on CSD in Appalachian English, suggesting that “apparent morphological influences are actually phonological influences that present themselves as morphological trends” (105).

TT attribute their finding of no morphological effect, which contrasts with all North American studies, to a difference between how CSD operates in British and American varieties. However, Smith et al. subsequently found the expected morphological effect for Scottish speakers, raising the possibility that TT’s finding might be restricted to only some British varieties, or simply spurious. Modeling CSD for the British speakers on Big Brother will provide another British data point.

Other linguistic factors A number of other linguistic factors influence the rate of CSD. Since deletion results in reduced pronunciations (i.e., a pronunciation where some segments are shortened or deleted), it might be expected that word frequency would be positively correlated with CSD rate, in line with the general finding that frequent words tend to be more reduced (e.g., Fidelholtz, 1975; Jurafsky et al., 2001; Schuchardt, 1885; Zipf, 1929). Several studies have argued for such effects on CSD (Bybee, 2000, 2007; Jurafsky et al., 2001; Myers and Guy, 1997). However, as pointed out by Walker (2012), these studies tend to examine frequency without controlling for other factors (e.g., phonological context, morphological class). Most studies where such factors have been considered alongside frequency have found no significant frequency effect (Hazen, 2011; Raymond et al., 2006; Walker, 2012), with the exception of Johnson (2012).[52]

The voicing status of the word-final consonant cluster also influences CSD. The standard finding is that deletion occurs more often for homo-voiced clusters (where the preceding consonant and the coronal are both voiceless or both voiced: *bend*, *fact*) than for hetero-voiced clusters (*bent*, *melt*) (e.g., Bayley, 1994; Fasold, 1972; Khan, 1991; Labov et al., 1968; Santa Ana, 1996; Wolfram, 1969). However, a strong effect in the opposite direction was found in Appalachian English (Hazen, 2011). Cluster voicing is not consid-

[52] There are some caveats related to methodological differences between the studies, which are discussed further when we examine the effect of frequency in the Big Brother data (Sec. 6.2.3)

ered by TT or Smith et al.

Non-linguistic factors Among properly social factors, ethnicity and age strongly affect CSD rate, while social class, sex, and other social factors have weak effects.[53] Ethnicity can be taken as a special case of differences between varieties, which in general show very different overall rates of CSD deletion (Schreier, 2005). Because of the small number and social heterogeneity of speakers in our dataset, we are not able to check for effects of social factors besides gender, with women expected to delete slightly more than men.

CSD rates tend to increase in more informal speaking styles (e.g., causal versus read speech) It is not clear whether this should be ascribed to speaking rate (informal speech tends to be faster), or a stylistic difference per se. Labov (2001: 196) argues that speakers have only mild social awareness of CSD, and show only moderate style shifting, at least for North American varieties. However, Smith et al. found large effects of situational context for the parents (but not the children) in their Scottish data. The most deletion was during play, followed by routine activities, followed by teaching and discipline. This effect is explained as decreased deletion rate in contexts where the parent is more likely to be monitoring her speech, and trying to "talk clearly" (90).

4.4.3 Vowel formants

Accents of English differ primarily in their vowels (Wells, 1982), making the dynamics of vowel quality in the Big Brother house of particular interest. Within a given variety of English, vowels differ in quality primarily in the first two formants (F1, F2), and secondarily in duration, F3, and other parameters. We model variation in F1 and F2 only.

Three vowels are considered. Two correspond to the *lexical sets* GOOSE and STRUT, using the notation of Wells (1982) for a vowel pronounced identically in a group of words

[53] See Hazen (2011) for an example, and Santa Ana (1992) for a review up to that point.

within a given dialect.[54] The third corresponds to the lexical set for TRAP, augmented for particular housemates to include all words which would have the same vowel as 'trap', for reasons discussed below; we call this modified variable TRAP′.[55] We will often use "GOOSE" rather than "the GOOSE vowel" to refer to the vowel shared by words in the GOOSE lexical set, and similarly for STRUT and TRAP′.

Variation in vowel formants is conditioned by many factors, most significantly characteristics of the speaker, and the linguistic context.

4.4.3.1 Speaker factors

Within a dialect region, the formant frequencies corresponding to each vowel vary greatly by speaker, due to physiological differences and social factors. Speaker gender has a particularly large effect; for example, women have higher pitch than men, and have larger vowel spaces (Simpson, 2009). However, significant variability remains between speakers of the same sex from the same dialect region, for a variety of reasons (Hillenbrand et al., 1995; Peterson and Barney, 1952). We will not explicitly model the effect of any by-speaker factors on vowel formants for the Big Brother data, due to data sparsity. However, we will model the (large) differences between speakers, which serve as a proxy.

How the vowel of each lexical set is realized (including F1 and F2) varies hugely across English dialect regions. Table 4.3 summarizes the expected pronunciations of GOOSE, STRUT, and TRAP′ in the dialect regions for each of the 12 speakers who are on the show for at least 50 episodes (we are not considering the others in the vowel analyses; see Sec. 5.2.4): Standard Southern British English (SSBE), Northern England, West Midlands, Southern Wales, Southern Scotland, General American, and General Australian.

[54] Wells defined each lexical set to represent a set of words whose vowels are pronounced identically in each of two reference accents–RP and General American—though in general the actual vowel quality used in each accent will differ. For example, three lexical sets are BATH, TRAP, and PALM. In GenAm 'bath' and 'trap' have the same vowel, which is different from the vowel in 'palm'. In RP 'bath' and 'palm' have the same vowel, which is different from the vowel in 'trap'.

[55] Note that this is no longer a lexical set, since it contains different words for different accents.

Table 4.3: Expected pronunciations of GOOSE, STRUT, and TRAP′ in the dialect regions represented by the 12 housemates present for at least 50 episodes. Sources: Beal (2004); Clark (2004); Cox and Palethorpe (2007); Ferragne and Pellegrino (2010); Penhallurick (2004); Stuart-Smith (2004); Watson (2007); Wells (1982).

	GOOSE	STRUT	TRAP′
SSBE	[uː]∼[ʉː]	[ʌ]∼[ɐ]	[æ]∼[a]
N. England	[uː]∼[ʉː]	[ʊ]	[a]
W. Midlands	[uː]∼[ʉː]	[ʊ]	[a]∼[æ]
S. Wales	[uː]∼[ʉː]	[ʌ]∼[ə]	[a]
S. Scotland	[ʉː]	[ʌ]	[a]
USA	[uː]∼[ʉː]	[ʌ]	[æ]
Australia	[ʉː]	[ɐ]	[æ]

GOOSE GOOSE varies between [uː] and [ʉː] in the dialect regions considered here. In some dialects (S. Scottish, Australian), [ʉː] has long been the norm. In many others, GOOSE is fronting (increased F2, little change in F1); this is common for young RP speakers (Hawkins and Midgley, 2005) and many American speakers (Labov et al., 2006). However, British dialects differ significantly in the extent of GOOSE-fronting, as reflected in a recent comparison of formant frequencies in 13 dialects (Ferragne and Pellegrino, 2010).

STRUT The primary dialect divide in England is between Northern English and Southern English varieties. The merger of STRUT with FOOT, usually as [ʊ], is one of the two most characteristic features of Northern English accents, and is strongly socially marked (Wells, 1982). In accents of the West Midlands (which lies on the isoglosses between Northern and Southern English accents), FOOT and STRUT are variably merged (Clark, 2004). The two vowels have not merged in other dialect regions represented in our dataset, and STRUT is realized as a lax vowel whose quality varies by dialect ([ʌ]∼[ɐ]∼[ə]).

TRAP′ In the US and Australia TRAP′ it is typically a clearly-fronted [æ]. In the UK, accents other than SSBE considered here tend to have the more central pronunciation [a]. Traditionally [æ] is used in SSBE, but Hawkins and Midgley (2005) found that younger

SSBE speakers have dramatically shifted towards [a]. Thus, the pronunciation of TRAP′ separates British and non-British accents in our dataset.

4.4.3.2 Linguistic factors

Consonantal context Vowel formants are strongly affected by coarticulation: the position of the articulators in producing a vowel is affected by nearby segments, which affects the resonant frequencies of the vocal tract, and hence the formant frequencies. The strongest coarticulatory effects come from the consonantal context (the identity of the preceding and following consonants). Consonantal context is the only linguistic factor controlled for in our models below, as an approximation of the total effect of coarticulation with surrounding segments. Consonantal context effects also differ somewhat by language and dialect. Two strands of studies, in the phonetic and sociolinguistic literatures, have examined the effects of adjacent consonants on vowel formants.

A sizable literature in articulatory phonetics details the effects of adjacent consonants on vowel formants in laboratory speech, in English and in other languages (see Steinlen, 2005). For English in particular, a seminal study by Stevens and House (1963) examined vowel formants in symmetric CVC environments for American English speakers, where C was one of 14 consonants (stops, fricatives, affricates) and V was one of eight vowels. In another important American English study, Hillenbrand et al. (2001) considered C_1VC_2 syllables (symmetric and asymmetric) for the same eight vowels as Stevens and House, with C_1 and C_2 one of the six stop consonants. Steinlen (2005) examined CVC syllables for British English speakers, where C was a stop consonant and V was one of 11 vowels. The general finding of these and other studies is that adjacent consonants strongly affect vowel formants, in complex ways which vary by consonant/vowel pair. An adjacent consonant's place tends to affect vowel formants more than its manner, and the effects of both manner and place differ by vowel quality, in particular for front versus back vowels. Following voiced consonants tend to lower F1 relative to following voiceless consonants, but do not have much effect on F2.

The effects of other consonant classes, such as nasals, laterals, and rhotics, have also been studied. For example, an adjacent nasal tends to strongly affect a vowel's formants, especially F1, because opening the velopharyngeal port introduces an extra pole-zero pair into the vocal tract transfer function in the vicinity of F1 (Stevens, 2000). Nasalized high/low vowels tend to have higher/lower F1 than their oral counterparts (Beddor, 1982).

Contextual effects on vowel formants have also been considered in the sociolinguistic literature. The study of sound change in progress using instrumental measurements was pioneered by Labov et al. (1972) in Philadelphia, and has resulted in a large literature on changes in progress in different speech communities (Labov, 1994, 2000). Such studies examine the 'internal' (linguistic) and 'external' (social) factors which condition variation and change in a particular (set of) vowel(s) in a community, and it is often the case that consonantal context is an important conditioning factors. Often it is the most important, as for GOOSE-fronting: across different dialects, GOOSE is most fronted before coronals and /j/, and least fronted before /l/ (e.g., Labov et al., 2006: Sec. 12.1; Fridland, 2008).

Other linguistic factors Vowel formants are affected by a host of other linguistic factors. Particularly important are coarticulation with adjacent vowels (see references in Beddor et al., 2002; Cole et al., 2010), syllable stress (e.g., Tiffany, 1959), and speaking rate (e.g., Gay, 1978). The overall shape of the vowel space is affected by stress and speaking rate, with vowels in unstressed syllables or faster speech tending to be centralized relative to vowels in stressed syllables or slower speech.

CHAPTER 5

BIG BROTHER: DATA

5.1 Show description

The corpus consists of speech from the eighth season of the reality television show Big Brother UK (Channel 4/Endemol), which aired from June 5 to September 5, 2008, for a total of 93 days. Before describing the show, two caveats related to the nature of the corpus merit discussion.

First, there is no reputable published source of information about the show or the housemates. Much of the information in this and the following sections is based on the extremely comprehensive Wikipedia pages for the 2008 season and for the broader franchise of Big Brother UK (Wikipedia, 2012a,b). In constructing both the speech corpus and a corpus of social interaction data (not discussed here) over two years, no one working on the project found a contradiction between the Wikipedia pages and the actual show, or anything housemates said. Nonetheless, the usual caveats about using Wikipedia as a reference source apply.

A second point relates to the passage of time on the show. We are using Big Brother as a "natural experiment" to study longitudinal variation. To do so, we measure phonetic variables in chunks of speech spoken at different times, and model the variables as a function of time and static factors. This procedure implies we trust the show's producers that speech broadcast on a certain day indeed comes from that day, and that speech represented as continuous is indeed continuous. We have no reason to doubt either one, a priori, especially since the show was broadcast every day. But we must bear in mind that either could sometimes be false, especially the latter: reality television shows are notorious for creative and subtle editing. We are also assuming that the speech broadcast from each speaker is linguistically random, in the sense that any time dependence observed in a housemate's use of a variable (after controlling for static factors) is not due to choices made by the producers about what chunks of speech to present.[1]

5.1.1 Show structure

The show was structured as follows. Sixteen contestants (known as "housemates") entered the house on Day 1, and five more were added at various points in the season. Housemates are gradually eliminated over the season, and the last one remaining wins a £100,000 cash prize. Most housemates leave because they are evicted, usually by a two-stage eviction process: each week, each housemate nominates two others for eviction, and the viewing public votes on which of the two who received the most votes should be evicted. There were a number of exceptions to this format during the show.[2]

During the season, housemates cannot leave the house, and do not interact with people not involved with the show. They do not have access to television, radio, books, music, or any other media, aside from when media are used as part of a house activity. Thus, for the duration of the show housemates have essentially no linguistic input from the outside world.[3]

Much of housemates' waking hours are spent interacting in some form with each other. (Even their sleep schedule is regulated, with sleeping during the day prohibited.) The main draw of the Big Brother franchise is the opportunity to observe such social interactions, both unstructured and structured. On the unstructured side, much of housemates' time is spent simply sitting around and talking. They cook, clean, and eat meals together. Many exercise regularly. On the structured side, housemates perform a series

[1] For example, suppose Rex (a housemate) produces vowel formants differently in the morning than in the evening, but otherwise does not change over time. If Rex's clips for the first half of the season were mostly from mornings, and later clips mostly from evenings, we might erroneously conclude he changed over the season.

[2] Five housemates remained on Day 93, and the winner was selected by the viewing public. Two housemates were evicted for threatening behavior (without a public vote), and one housemate chose to leave. Individual housemates were also sometimes automatically nominated or stripped of nominating rights during a given week, for various reasons.

[3] On a few occasions Big Brother provided housemates with letters or videos from family members as a reward for some task.

of tasks throughout the season, either individually or in teams. Big Brother provides rewards and punishments based on performance in tasks.

Housemates can also go to the *diary room*, a room isolated from the rest of the house, to speak with Big Brother. Big Brother is in fact different people in different clips, both male and female, with a broad range of accents. Big Brother can see housemates, but cannot be seen by them, and communicates only through audio. Housemates go to the diary room to talk about events in the house, to vent about other housemates, to speak about their feelings, to make requests, and so on. They can be called to the diary room to nominate other housemates, to receive a message, for disciplinary purposes, or for Big Brother to ask their opinion on events in the house or on other housemates. For reasons discussed below, the corpus used in this study consists exclusively of speech from the diary room.

5.1.2 Housemates' audio and video data

Housemates are recorded at all times in the house, both by cameras and microphones present in most rooms, and by wearable microphones which housemates nearly always have on. It is not clear how much of the audio broadcast during the show comes from each of the two types of microphone, or what post-processing has been performed. But in general the sound quality is very good, especially in the diary room.

During the season, the public could see recordings of housemates via a live feed or daily produced episodes. The pay TV channel E4 showed a live feed for most of the day for the duration of the season, consisting of footage from one camera at a time, broadcast with a time delay to censor conversation about certain topics (e.g., persons outside the house).[4] The live feed did not show footage from the diary room. A daily one-hour "highlights show" was shown on the public channel Channel 4. These produced episodes consisted of continuous segments (usually 1-10 minutes) from the previous day in any room of the house (including the diary room), presented without commentary.

[4] The live feed was cut daily during the broadcast of "E4 Music", but it is not clear for how long.

This broadcast setup provides the context for every utterance by a housemate. As Thornborrow and Morris (2004: 248) note, unless in the diary room, "for the inmates of the Big Brother house, 'liveness' provides a permanent context for any interaction that takes place between them. Not all of their talk is broadcast live, of course, but any of it *could* be." In the diary room, a housemate knows that anything she says could be broadcast in the day's produced episode.

Given this context, some caution is warranted about the ecological validity of housemates' speech. We are interested in Big Brother as a natural experiment where we can observe longitudinal variation within individuals. But how much our findings bear on linguistic variation in the real world depends on a crucial question: to what extent are housemates "acting naturally" and producing spontaneous speech, versus producing monitored speech as part of a performance? Because housemates are constantly observed, data from Big Brother entails an extreme version of Labov's *observer paradox*: "to obtain the data most important for linguistic theory, we have to observe how people speak when they are not being observed" (Labov, 1972: 113).

What does this mean for housemates' speech? On the one hand, housemates are aware that they are being observed, and anything they say may be broadcast on television. On the other hand, housemates also know that little of what they say will *actually* be broadcast on the live feed, and even less on a produced episode. It must also be kept in mind that the Big Brother house becomes housemates' world for three months. From our perspective, the hope is that housemates will forget they are being observed at least enough to produce the same register of speech that they would outside the house.

Certainly, the overwhelming impression is that housemates' speech is spontaneous and conversational. It shows many characteristics of conversational speech, like extensive segmental reduction. But there are certain situations where housemates can be much more sure they will be broadcast—such as when nominating other housemates for eviction, or taking part in tasks—and their speech in these situations sometimes seems more

Table 5.1: Information on the 21 housemates of Big Brother 9 UK. "HM speech" refers to the amount of speech from this housemate in the corpus. Housemates are listed in reverse order of the day they left the house (winner first).

Housemate	Sex	Age	Origin	Country/Region	Days	DR clips	HM speech
Rachel	F	24	Torfaen	Wales	1–93	40	30:10
Michael	M	34	North Ayreshire	S. Scotland	1–93	49	58:17
Sara	F	27	Melbourne	Australia	30–93	20	20:23
Rex	M	24	London	London	1–93	39	44:11
Darnell	M	26	Missouri	USA	1–93	42	46:56
Kathreya	F	31	??	Thailand	1–90	26	24:17
Mohamed	M	25	??	London/Somalia	1–90	35	30:41
Lisa	F	40	Cheshire	N. England	1–86	34	35:46
Nicole	F	19	Surrey	S. England	58–79	12	13:13
Stuart	M	25	Greater Manchester	N. England	16–72	18	19:58
Dale	M	21	Merseyside	N. England	1–65	29	23:17
Luke	M	21	Greater Manchester	N. England	1–58	28	39:03
Maysoon	F	??	London	London	30–56	4	2:34
Rebecca	F	21	West Midlands	Midlands	1–51	17	14:40
Belinda	F	44	Devon	S. England	30–44	4	4:56
Mario	M	43	Cheshire	N. England	1–37	8	8:08
Jennifer	F	22	County Durham	N. England	1–30	8	11:41
Sylvia	F	22	??	??/Sierra Leone	1–23	3	4:58
Dennis	M	24	Edinburgh	S. Scotland	1–23	7	4:36
Alexandra	F	23	Greater London	London	1–14	11	17:46
Stephanie	F	19	Liverpool	N. England	1–9	5	3:23
Total						439	06:39:05

monitored. The speech considered here all comes from the diary room, which may lead to it being more or less monitored. Diary room speech is often emotional, and emotional speech tends to be more spontaneous. Also, housemates may speak more spontaneously in the knowledge that diary room speech cannot be broadcast to the live feed. On the other hand, they know it can be broadcast in a produced episode, and often is. In sum, legitimate questions exist about the extent to which speech from Big Brother is "real"—as is also true to some degree for any other situation where speakers are being knowingly recorded (e.g., sociolinguistic interviews, ethnographic studies). We will consider housemates' speech in the diary room to be at least spontaneous, and arguably conversational.

5.1.3 Housemates

Table 5.1 gives information about each of the 21 housemates who took part in the show: name, place(s) of origin, the length of time on the show, and amount of data in the corpus (both in number of clips, and in amount of speech by the housemate). Knowing each housemate's personal background is important for understanding their accent. Unfortunately, little information on housemates' biographies is available beyond what is listed online; however, this information at least seems to be accurate (see Sec. 5.1).

The housemates have very different accents, corresponding to their diverse geographic origins and social backgrounds. Sixteen housemates are native speakers of British dialects, which seem to broadly correspond to their home regions.[5] Belinda and Nicole come from southern England, and have essentially Standard Southern British English (SSBE) accents. Alexandra, Rex, and Maysoon come from London, but have relatively different accents: Rex sounds the most 'posh' (closest to SSBE), while Alexandra has many features associated with working-class speakers. Seven speakers come from northern England, and all have accents which are immediately identifiable as Northern (e.g., merger of /ʊ/ and /ʌ/, distinctive prosodic patterns). Rebecca comes from the West Midlands (Coventry) and has some features characteristic of this region. Dennis and Michael come from southern Scotland, and have strong regional accents (Edinburgh, Ayrshire). Rachel comes from southern Wales, and has a mild regional accent (e.g., [ø] for NURSE).

The backgrounds of the other housemates are less straightforward. Sara is Australian and has a strong Australian accent, but has lived in London for several years. Darnell was born in the UK, but was raised in the US (St. Louis), then was deported to the UK as an adult. His accent is clearly American, with some minor African American Vernacular English features (mostly in lexical choice). (He is African American and albinistic.) The remaining three housemates have perceptibly non-native accents. Sylvia was born in

[5] Note that although I have some familiarity with British dialectology, I do not command a British dialect. The following remarks are my best guesses based on my knowledge of British dialects, and some consultation with (non-linguist) British speakers.

Sierra Leone, and moved to the UK at age 11; Mohamed was born in Somalia, and moved to the UK with his family (age unknown). Both speak near-native English with relatively light accents. Kathreya is a native Thai speaker who moved to the UK as an adult. Her English is heavily accented and frequently ungrammatical. She is the only housemate with a clearly non-native command of English.

5.2 Corpus description

The corpus consists of all segments from the produced episodes where a single housemate is in the diary room, which we call *diary room clips*, or simply *clips*. There are 439 clips in all, corresponding to a total of 6 hours and 39 minutes of housemate speech.[6] The number of clips and amount of speech for each housemate are shown in Table 5.1. The corpus is highly unbalanced, with 30–60 minutes of speech for some housemates, and only a few minutes for others. This is largely because different housemates are on the show for different periods of time, but also because diary room clips for some housemates occur more frequently than others. For example, Luke is on the show for only 60 days, but there is more speech for him than for some housemates who were on the show for three months. Audio quality for the diary room clips is generally very good, with the exception of some background noise (often from housemates moving around), and there were very few problems with performing phonetic measurements from spectra and waveforms.

The corpus is limited to diary room clips of single housemates for both practical and scientific reasons. It was not feasible to transcribe more than a fraction of the speech contained in the produced episodes (93 hours) given available resources, so we needed to somehow limit the amount of speech considered. Since the primary goal of the study was to examine longitudinal variation in individuals, we decided to consider only one type of speech, of the many which occur in the corpus (conversations between housemates,

[6] The actual total length of the clips is significantly longer, owing to speech from Big Brother, footage of events outside the diary room, etc.

speech during tasks, diary room speech, etc.), so that the type of interaction a housemate was engaged in at different time points would be controlled for. Diary room speech was a natural choice, since the diary room provides a constant recording environment and a relatively constant social setting (compared to speech outside the diary room). Finally, we decided *not* to consider diary room clips containing interactions between housemates. Because the only interaction in the clips is with Big Brother, whose role is usually limited to brief questions or answers, the speech in the corpus is roughly comparable to the type of spontaneous speech commonly elicited in sociolinguistic interviews. The corpus allows us to examine how each housemate's 'baseline' linguistic usage varies over time, abstracting away from short-term shifts which may occur during conversation. This also means that the corpus does not necessarily say anything about any changes in housemates' speech *when they are talking with each other*. Exactly what can be concluded from the corpus about longitudinal variation in individuals is discussed when we build dynamic models in Chapter 7.

5.2.1 Transcription

Each clip was orthographically transcribed in three stages. A preliminary transcription was performed by seven undergraduate research assistants (three at the University of Chicago, four at the Massachusetts Institute of Technology), none of whom had prior experience with this corpus or with British dialects.[7] Two of these RAs, Maria Nelson and Natalie Rothfels, continued working on the project and became familiar with each housemate's speech. They performed a second pass, in which the original transcriptions were corrected and made consistent. Finally, I reviewed and corrected all transcriptions.

Priority was given in transcription to representing as faithfully as possible what was said in each segment of speech, with an eye towards forced alignment (Sec. 5.2.4.1), where

[7] The Chicago transcribers were supervised by myself, and the MIT transcribers were supervised by Peter Graff.

the transcription would be transformed into a string of phones and aligned with the signal. Non-standard usage or lexical items were included, and non-lexical fragments of speech were transcribed phonetically. Non-speech noises were segmented out from segments of speech, so that each transcribed interval consisted only of speech and silence.

5.2.2 Voice onset time

The VOT dataset consists of three types of annotation for each stop: a manual VOT measurement, an automatic VOT measurement, and the values of predictors to be used in the static and dynamic models.

Measurements A total of 6494 manual VOT measurements were performed by three transcribers. The annotation procedure is described in Appendix A.1. Nineteen tokens where it was unclear how to apply the measurement criteria were excluded.

The algorithm described in Chapter 3 was used to assign an automatic measurement for each stop,. Recall that to produce a VOT measurement for a given stop, the algorithm needs a classifier which has been trained on manually-labeled data, and the left boundary for the stop's host word, which specifies where the algorithm should begin looking for the burst. The classifier trained in the BB base experiment (Sec. 3.5.1) was used, with all the algorithm's parameters set as in that experiment. For each stop, 50 ms before the *manually-annotated* burst onset was taken to be the left word boundary. Note that, like in all experiments presented in Chapter 3, we are taking the approximate location of the left word boundary as given, rather than itself determined by an automatic procedure (such as forced alignment to an orthographic transcript).

Predictors Information about following segments was determined from CELEX (Baayen et al., 1996). The word associated with each token was defined orthographically (i.e., the noun and verb forms of *wind* would be equivalent). The frequency of a word was defined as the frequency of the highest-frequency wordform in CELEX with the same spelling. A word's initial syllable was considered to be stressed if it was listed as having primary stress in the CELEX primary pronunciation, and was not a function word: *to*, *'cause*, *can*, *can't*, *could*, *couldn't*, *'til*, *'kay*.

Speaking rate was determined by a more involved process. Each transcribed chunk of speech was force-aligned with a sequence of phones corresponding to its canonical pronunciation (see Sec. 5.2.4.1). The forced aligner also inserts silences for regions that are not associated with a segment. Thus, after forced alignment, we have a predicted beginning and end for each word in the transcription, and the locations of all silences. A *spurt* of speech was defined as a sequence of words separated by a silence of 60 msec or greater. The speaking rate within a spurt was defined as the number of syllables in the canonical realization of its transcription, divided by the spurt's duration. The speaking rate predictor for each token was the rate for the spurt containing the VOT's host word.

Outliers Outliers of two types were excluded. Tokens with extreme values of the dependent variable (log(VOT), in our models) within a given speaker are likely to be overly influential when fitting models. Accordingly, for each speaker, all tokens whose manually-measured log(VOT) was greater than three standard deviations from the mean were excluded when building models of manually-measured VOTs. Tokens with manually-measured VOT below 10 msec or above 250 msec were also excluded, as extreme values across all speakers; these cutoffs were chosen by visual inspection of the distribution of VOTs. An analogous procedure was followed for automatically-measured VOTs when building models of automatically-measured VOTs. In total, 111 tokens (1.7%) were excluded for manual measurements and 85 tokens (1.3%) for automatic measurements.

Next, inspection of the speaking rate predictor showed that many speakers had a small fraction of tokens with extreme rate values, relative to other tokens. Because speaking rate was determined using a semi-automatic process (see above), it was thought that speaking

Table 5.2: Descriptive statistics for each housemate's VOT data. N and N_{clip} denote the number of data points and the number of clips per housemate. "Mean VOT" is the average of each speaker's mean VOT value for each word, in msec.

Speaker	N	N_{clip}	Mean VOT
Stephanie	41	5	64.6
Maysoon	45	4	58.4
Sylvia	48	3	76.1
Belinda	74	4	63.0
Dennis	85	4	59.8
Mario	110	8	67.8
Rebecca	141	16	65.1
Stuart	185	17	72.0
Nicole	195	11	59.6
Jennifer	205	8	60.3
Sara	209	16	55.5
Kathreya	287	24	75.2
Dale	342	23	72.8
Rachel	343	34	68.7
Alexandra	348	10	61.3
Mohamed	380	28	43.8
Lisa	451	31	67.4
Darnell	536	36	51.0
Rex	597	38	55.4
Luke	632	28	41.1
Michael	786	44	68.9

rate outliers were largely due to errors in this process. Because speaking rate has a large effect on VOT compared to other predictors, it was decided to exclude tokens where the speaking rate predictor was probably in error, by the same criteria as used for log(VOT).

The final dataset consisted of 6494 tokens from nearly all clips (398 total) for all 21 housemates.[8] Table 5.2 shows the distribution of tokens and clips among housemates.

[8] A few clips were excluded whose annotation files were lost.

Table 5.3: Descriptive statistics for each housemate's CSD data. Asterisks denote British housemates. N and N_{clip} denote the number of data points and the number of clips per housemate. "Deletion rate" is the mean of the proportion of deleted tokens for all words used by a housemate.

Housemate	N	N_{clip}	Deletion rate
Michael*	777	41	0.41
Darnell	737	39	0.50
Rex*	587	28	0.51
Luke*	568	28	0.36
Lisa*	493	30	0.40
Rachel*	487	38	0.41
Nicole*	297	11	0.61
Dale*	293	24	0.32
Mohamed	289	24	0.60
Alexandra*	284	9	0.32
Sara	263	16	0.64
Stuart*	262	17	0.33
Jennifer*	261	8	0.52
Rebecca*	231	16	0.48
Mario*	117	8	0.42
Dennis*	98	5	0.37
Belinda*	64	4	0.43

5.2.3 *Coronal stop deletion*

The CSD dataset consists of three types of annotations for words in the corpus ending in t/d-final consonant clusters: the host word's CELEX wordform ID, the realization of the final coronal, and the phonological context surrounding the final coronal. The annotation process is described in Appendix A.2. We annotated every word ending in a t/d-final cluster (with some exceptions, described below) in nearly all clips (388 total) from 17 of the housemates, resulting in 6108 tokens.[9] Table 5.3 shows the distribution of data points and clips among housemates.

In studies of CSD, a number of processing steps are usually applied to the data before

[9] A few clips were excluded whose annotation files were lost. Two housemates (corresponding to 8 clips) were not coded. One housemate who deletes near-categorically (Kathreya) was excluded. Kathreya has a heavy Thai accent, and seems to produce very few word-final CC clusters. Because she behaves so differently from other speakers, including her has an undue impact on the static models of CSD.

building a statistical model. Two such steps are taken here, while others are not.

Steps taken The vast majority of studies on CSD treat realization as a binary variable: the final coronal is present or not. This simplification belies the many ways a word-final coronal can be realized, for example as a burst, a glottal stop, or completely absent. Recent studies have begun to examine different surface realizations of /t/, with particular attention to their social meanings., in both American and British varieties (e.g., Foulkes et al., 2005; Levon, 2006; Podesva, 2006). However, there are two important reasons to stick with a binary division. First, comparing results with most previous work (where the binary division was used) is much easier. Second, modeling grouped data turns out to be significantly harder for a multinomial response than for a binomial response. Thus, we leave the exciting prospect of building a mixed-effects multinomial model of final coronal realization to future work, and dichotomize the realization annotation into a binary response: absent versus present.[10] Our dichotomization follows TT in treating any unambiguous phonetic reflex of the underlying coronal as being its surface realization.[11]

We also follow TT in discarding all tokens whose final cluster is /rt/ or /rd/ in American English. Many housemates are speakers of non-rhotic varieties of English, where /r/ does not occur in codas (e.g., in RP: *start* [stɑːt], *dear* [dɪə]), while others are speakers of rhotic varieties (e.g., Australian, American English). Thus, for some housemates words ending in /rt/ or /rd/ are subject to CSD, while for others they are not. Including these words in the model would be problematic, because the predictors for preceding phonological context would have different meanings for the two groups of speakers.

Steps not taken The practice in most studies of CSD is to discard several other types of tokens which are not discarded here.

[10] In terms of the labels we used in realization annotation (Appendix A.2), `NONE`, `NONE_BUT`, `SHARED_COR`, and `SHARED_IDF` counted as "absent", and all other labels as "present".

[11] Note that TT discard all tokens before interdental fricatives or coronals, so our dichotomization is slightly different.

Some high-frequency words (especially *and*) are often excluded, because their deletion rates can be near-categorical. Tokens in the "neutralizing contexts" of a following alveolar consonant (/t/, /d/, sometimes also /s/ or /n/), where deletion rates can be very high, are also typically excluded. The rationale in both cases is that CSD is so frequent as to no longer be a variable process; also, fitting logistic regression models with near-categorical cases can lead to problems with model identifiability (Guy, 1988; Tagliamonte, 2006: 86–87). These cases show high, but not near-categorical deletion rates in the Big Brother data. Our approach is to build a statistical model which allows for deletion rates specific to particular words (via a by-word random intercept) and for a strong effect of "neutralizing contexts" on deletion rate (via a fixed-effect predictor specifying whether the following context is neutralizing), rather than discarding certain words and phonological contexts a priori.[12]

It is also common practice to perform type/token sampling, where only a fixed number of data points for a given word per speaker are included, to restrict the potential influence of any one lexical item (see Tagliamonte, 2006: 95–96). This concern is justified when performing a classic logistic regression, where only fixed effects are modeled, as in the programs (VARBRUL or GOLDVARB) used to build logistic regression models in most CSD studies. Since the grouping of the data points by word is not modeled, type/token sampling is carried out so that the data is balanced, and each word does not have an influence on the model proportional to its frequency.[13] This step is not necessary for the mixed-effects models used here, where the grouping of data points by lexical item is included in the model (as a random intercept). In general, mixed models are robust to "unbalanced" data, where the distribution of tokens among different units in a grouping level (e.g., different words) is very uneven.

[12] To give an idea of the impact of leaving this data in, 41% of tokens in our dataset either are from the word *and* or are followed by a word beginning with /t/ or /d/.

[13] As noted by Tagliamonte and Baayen (2012), even this practice is potentially problematic: the non-independence of data points for the same word is not being modeled, violating the independence assumption of classic logistic regression.

5.2.4 Vowel formants

The dataset consists only of tokens of GOOSE, STRUT, and TRAP from the 12 speakers who were in the Big Brother house for 50 days or longer. These are the only speakers who will be included in the dynamic analyses (for all variables) in the next chapter, essentially because there is too little data for other speakers to assess time dependence. Also, unlike the static models for VOT and coronal stop deletion, the static models for vowel formants will serve largely as a preliminary step for the dynamic analyses. Thus, there is little point to including speakers besides those included in the dynamic analyses.

We determined which vowels to consider as data from the GOOSE, STRUT, and TRAP classes as follows. The automatic formant measurement suite described below uses CMUDict, a widely-used electronic dictionary of American English;[14] we thus used CMUDict as the starting point for determining which words had vowels falling into a particular class, after excluding non-lexical items (e.g., fillers, exclamations: *um*, *uh*, *woo*).

For GOOSE, all vowels marked as `UW` in CMUDict were included. For STRUT, all marked as stressed `AH` in CMUDict were included, with the exception of vowels which can be pronounced as both STRUT and LOT in British varieties.[15]

For TRAP′, some explanation is merited. We initially wanted to consider the BATH lexical class, arguably the most socially salient in British dialects (Wells, 1982). BATH consists of a few dozen words which are pronounced like TRAP in American English and like PALM in SSBE. Unfortunately, because so few words are in this class, there turned out not to be enough tokens in our dataset to examine longitudinal variation within individuals. We opted instead to consider TRAP, which at least varies in quality between accents. Even then, there were not quite enough tokens to examine longitudinal variation for many individuals. So we finally chose to examine all vowels which were very likely to be pronounced like 'trap' *in a particular speaker's dialect*; this is the set of vowels

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

[15] The STRUT vowel in *what*, *was*, *from*, *of*, *hovel*, *somebody*, *anybody*, *nobody*, *everybody*.

we are calling TRAP′. We used a conservative procedure, using the lists of BATH words given by Wells (1982) and in a recent educational web site by the British Library (2007), to determine which words beyond TRAP were in TRAP′ for a particular speaker.[16]

5.2.4.1 Formant measurement

For each token in the GOOSE, STRUT or TRAP′ classes, we obtained F1 and F2 measurements using an automatic method, followed by several post-processing steps.[17]

Automatic formant measurement was performed using the FAVE program suite developed at the University of Pennsylvania (Rosenfelder et al., 2011). FAVE takes as input audio files and their orthographic transcriptions, and works in two steps.

First, each transcribed file is aligned with a sequence of phones corresponding to its canonical pronunciation in a reference dictionary, a procedure called *forced alignment*. FAVE uses a HMM-based aligner built using the HTK toolkit,[18] with acoustic models (probabilistic models of how each phone can be pronounced) for the American English phone set from CMUDict, trained on American English speech. CMUDict is used as the reference dictionary to determine possible pronunciations for each word in an orthographic transcription. We augmented CMUDict with American English pronunciations for all words in our corpus that are not in CMUDict, as well as pronunciations for all segments of speech which were not full words (false starts, etc.); we then aligned each DR clip in our corpus with its transcription, resulting in a predicted start and end time for each phone in the sequence of phones implied by the transcription.[19]

[16] That is, we attempted to minimize words marked TRAP′ which in fact had the PALM vowel.

[17] This process was carried out by myself and Rachel Hwang, an undergraduate research assistant.

[18] `http://htk.eng.cam.ac.uk/`

[19] Note that we are using an aligner trained on American English speech, and that uses an American English dictionary. This mismatch turns out to not be problematic for our purposes, since the aligner still determines vowel boundaries accurately, and there is usually a one-to-one correspondence between vowels in British and American varieties (with the exception of vowels followed by /r/ codas in American English).

Next, automatic formant measurement is performed for all vowels in each audio file, using the predicted start and end time for each vowel from forced alignment. FAVE includes a range of configuration options for automatic measurements. Under the options we used, measurement is performed as follows. For a given audio file, the LPC-based Burg method in Praat (Boersma and Weenink, 2011) is used to compute formant tracks corresponding to 3, 4, 5, and 6 formants, for the whole file. To avoid measuring reduced vowels, only vowels with duration of at least 50 msec are considered. For each vowel, choosing values for F1, F2, and F3 requires choosing at what point in the vowel to sample the formant tracks, and which format track to associate with which formant. First, a measurement point is chosen using the `faav` method.[20] The question is now which number of formants to use, and which formant tracks correspond to F1, F2, and F3. Using the method of Evanini et al. (2009), the choice of F1, F2, and F3 is made that minimizes the Mahalanobis distance from the distribution of the vowel's label, using a multivariate normal distribution based on all instances of that vowel in the Atlas of North American English (Labov et al., 2006). Finally, the distribution for each vowel label is re-computed based on the speaker's estimated formant values for all instances of that label in the audio file, and the formant values are re-estimated using the minimum-distance method, but now using the re-computed distributions.

Reduced vowel exclusion and outlier correction When studying vowel formants, tokens that are clearly reduced (i.e., neutralized to [ə]) are normally not of interest; thus, a common practice when manually measuring formants is to exclude heavily-reduced tokens. When automatically measuring formants, it is not known a priori whether each token is reduced or not. We took several steps to avoid reduced tokens. The first, excluding vowels with duration of <50ms, was mentioned above. Because reduction is more likely

[20] Formants are measured at one-third of the vowel's duration, with a few exceptions: PRICE and FACE are measured at the maximum F1 value (on the F1 track for the vowel), GOAT and MOUTH halfway between the maximum F1 value and the beginning of the vowel, and GOOSE is measured at the beginning of the vowel when preceded by a coronal consonant.

for (phonologically) short vowels, unstressed vowels, and high-frequency words, we also excluded all unstressed tokens (according to the CMUDict transcription), and for the two short vowel variables STRUT and TRAP′, we excluded tokens from a list of high-frequency words.[21] These steps make for a highly-conservative procedure, which is partially responsible for the relatively small number of tokens per vowel in the final dataset. Future work will refine this procedure, to reduce the number of unreduced discarded tokens.

Finally, a subset of the automatic measurements were manually corrected. Some highly inaccurate measurements remained despite the involved forced alignment and formant measurement procedures, usually because the wrong formant tracks were chosen as F1, F2, and F3.[22] For each vowel for each speaker, we visually determined possible mismeasured tokens by plotting all tokens with a superimposed 80% confidence ellipse. We were conservative (i.e., selecting many tokens that were probably correctly measured) in selecting tokens for re-measurement, and discarded all tokens where the suspect formant values were due to very high pitch, screaming, a reduced or deleted vowel, or speech which was otherwise unusual. For other tokens we manually re-measured formants only when the automatic measurements were at least 50–100 Hz off.

Vowel normalization Finally, because we are interested in comparing formants across different speakers, we performed vowel normalization to control for physiological differences. Adank et al. (2004) find that the best-performing vowel normalization procedure (in terms of a discriminant analysis of Dutch vowels following normalization) is *Lobanov normalization* (Lobanov, 1971); each speaker′s j^{th} formant is scaled by subtracting its mean and dividing by its standard deviation (both across all the speaker′s tokens). We always use Lobanov-normalized formants below, written simply as F1 and F2.

[21] Namely: *the, but, of, to, do a, am, an, and, as, at, has, have, nah, than, that, that′s*.

[22] Secondarily, because of an incorrect measurement point, inaccurate force-aligned phone boundaries, or reduced tokens which were not caught.

Table 5.4: Number of tokens for each housemate for each of the three vocalic variables.

Housemate	GOOSE	STRUT	TRAP′
Dale	79	90	57
Darnell	207	248	273
Kathreya	112	164	95
Lisa	136	192	150
Luke	186	220	204
Michael	193	379	337
Mohamed	108	123	57
Rachel	135	165	95
Rebecca	31	36	70
Rex	247	263	123
Sara	50	128	113
Stuart	58	74	64
Total	1542	2082	1638

5.2.4.2 Empirical data

The final dataset contains 1542 tokens for GOOSE, 2082 tokens for STRUT, and 1638 tokens for TRAP′. Table 5.4 shows the distribution of tokens among the 12 housemates. As a quality check of the formant data, we verify that speakers' formants for each vowel pattern as expected, given their dialect regions (see Sec. 4.4.3.1).

GOOSE Fig. 5.1 shows F1/F2 plots for GOOSE. The one clearly L2 speaker, Kathreya, has a proper high back rounded vowel, which is characteristic of her L1 (Thai; Tingsabadh and Abramson, 1993). All other speakers show various degrees of fronting, as expected given that GOOSE-fronting is occurring in many varieties of English. GOOSE is particularly fronted for Michael, Dale, and Rex. Michael is from Scotland, where the fronted variant is standard.[23] Dale is from near Liverpool, and Rex is a young speaker of SSBE; recent work suggests that the fronted variant is expected for both (Ferragne and Pellegrino, 2010; Hawkins and Midgley, 2005).

[23] However, it is also standard in Australia, where Sara is from, and she shows somewhat less fronting.

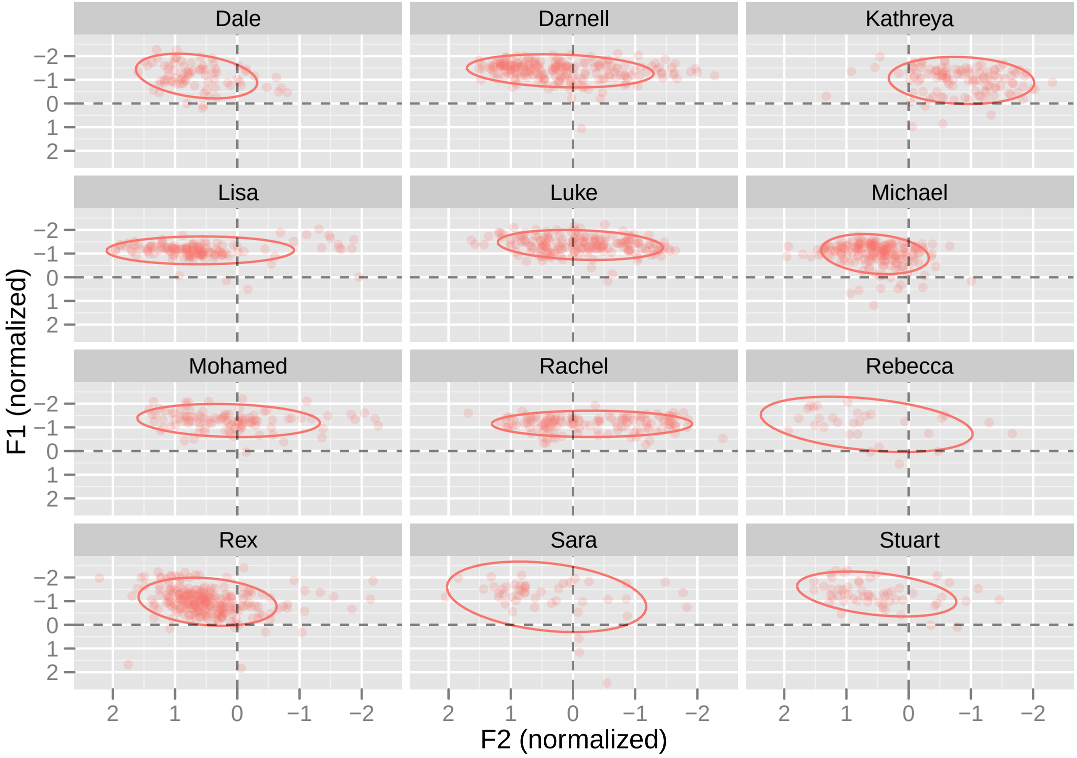

Figure 5.1: Lobanov-normalized F1 and F2 for GOOSE (red dots) vowel tokens for each speaker, with ellipses containing 80% of tokens.

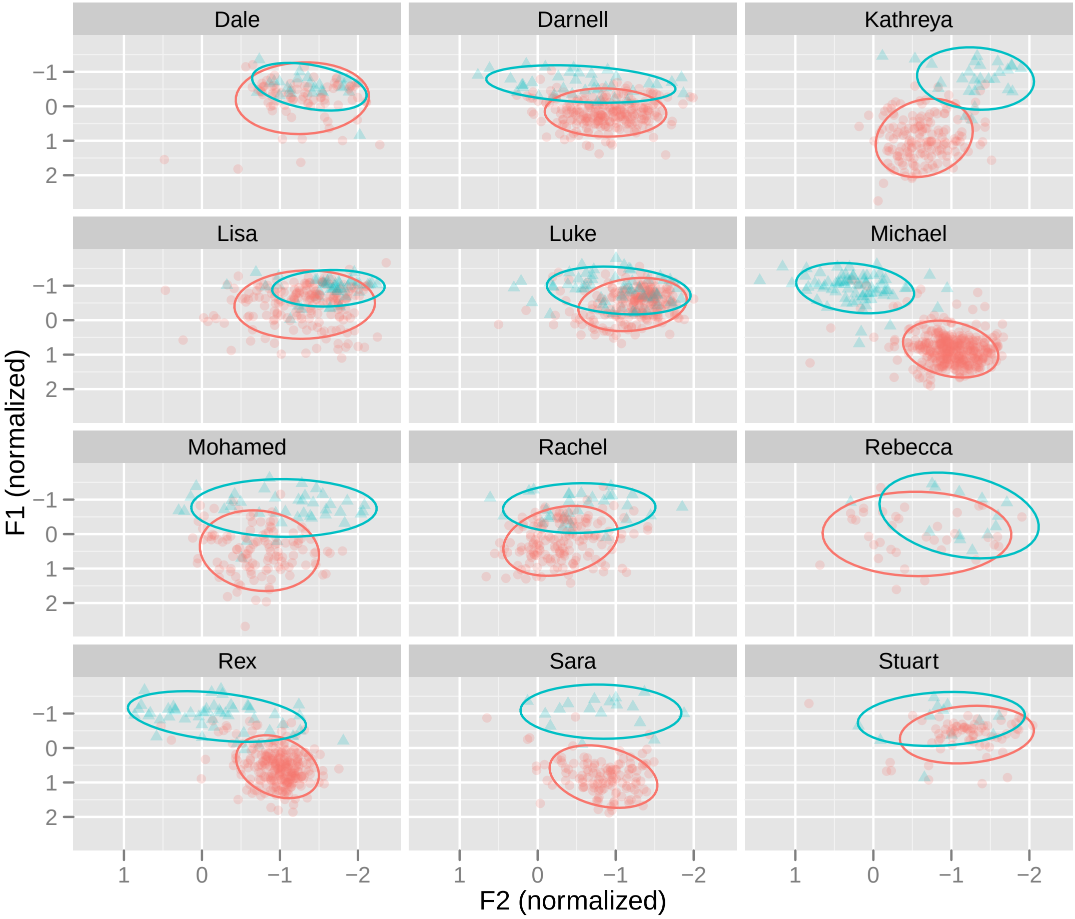

Figure 5.2: Lobanov-normalized F1 and F2 for FOOT (blue triangles) and STRUT (red dots) vowel tokens for each speaker, with ellipses containing 80% of tokens.

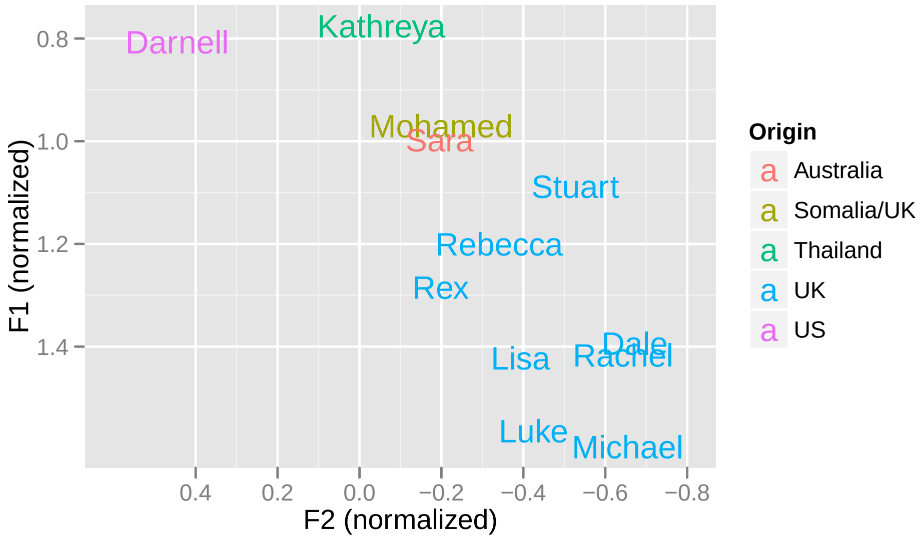

Figure 5.3: Mean Lobanov-normalized F1 and F2 for each speaker for TRAP′. Color indicates speaker's country of origin. Housemates range from [a] (Michael) to [æ] (Darnell).

STRUT Fig. 5.2 shows each speaker's tokens for STRUT, with tokens for FOOT plotted as well to show the amount of overlap between the two. Recall that the FOOT/STRUT merger is a paradigmatic feature of Northern English accents. Among the 12 housemates, Dale, Lisa, Luke, and Stuart are from Northern England. They are the housemates with the greatest overlap between the two categories, as expected. Rebecca also shows some overlap, as expected from her dialect region (West Midlands), which overlaps with the isogloss for the merger.

TRAP′ For TRAP′, differences between housemates are best visualized by plotting mean TRAP′ realizations in the F1/F2 plane (Fig. 5.3). Housemates vary roughly along a line from [æ] (upper left) to [a] (lower right). There is a clear division between housemates with native British accents and housemates with other accents, as expected from Table 4.3: British speakers are closer to [a]. Darnell has a clear [æ], as expected for a General American accent. With one exception (Stuart), British speakers from regions where [a] is expected (Luke, Michael, Rachel, Lisa, Dale, Stuart) are closer to [a] than British speakers from regions where both [æ] and [a] are used (Rex, Rebecca).

CHAPTER 6

BIG BROTHER: STATIC MODELS

We now turn to modeling phonetic variation in the Big Brother house. This chapter develops *static models* of synchronic variation for each of the five variables, without taking time into account, and discusses their results with respect to previous work. The next chapter describes *dynamic models* of longitudinal variation in each variable within individuals.

Motivation A first motivation for building static models is as an important preliminary step for the dynamic models. There is tremendous synchronic variation in how segments are realized at a given time by an individual, as a function of static factors (linguistic context, social factors, etc.) Each of our phonetic variables is a property of individual segments. Thus, there are many reasons why a variable might take on different values at different times, simply because of what is being said and who is saying it.[1] We are interested in what time dependence remains in a housemate's use of a variable *after* controlling for static factors. Modeling synchronic variation in the variable across all speakers is a logical first step, to understand what static factors need to be controlled for.

A second motivation is that the static models are interesting case studies of phonetic variation in spontaneous speech in their own right. The models for VOT and CSD in particular yield interesting and surprising findings with respect to previous work.

Most of what is known about VOT comes from laboratory studies of planned speech, where the effects of a small set of factors are precisely determined, with all others held constant (see Sec. 4.4.1). Studying VOT variation in spontaneous speech corpora is important for understanding to what extent the results of these studies carry over to speech

[1] c.f. Prince (1987: 83), in an early study of dialect shift in individuals: "[dialect shift] turns out to be amenable to study, as a special case of stylistic variation. That is, dialect shift may be construed as a change in the pattern of stylistic variation over time, for each relevant variable, where normal factors affecting such variation are controlled."

in the wild.[2] In addition, corpus studies of VOT are needed to examine the relative importance of the conditioning factors found in laboratory studies, how much these factors vary by speaker, and how much of the VOT variation in continuous speech is predictable. Two small previous corpus studies reached intriguing conclusions relative to laboratory studies on these points, suggesting that more are needed (see Sec. 4.4.1).

Most studies of CSD have examined North American varieties (see Sec. 4.4.2). In particular, there have been two studies of British speakers, by Tagliamonte and Temple (2005) and Smith et al. (2009), abbreviated TT and Smith et al. One contribution of our static models of CSD is simply to contribute another study of British speakers, albeit in our case from very different dialect regions. We also replicate a surprising finding of TT on morphological conditioning of CSD which conflicts with studies on North American varieties. Finally, a number of methodological points come up that are relevant to the interpretation of previous work, and for the design of future studies.

6.1 Voice onset time

Two models were built for the dataset described in Sec. 5.2.2, one using manually-measured VOTs and one using automatically-measured VOTs, which we call the *manual model* and the *automatic model*. The models were fit to the *manual dataset* and *automatic dataset*, which (besides the type of measurement) differ only in which tokens were excluded as outliers.

In each model, the (natural) logarithm of VOT was modeled using a linear mixed-effects regression model with random effects for speakers and words, of the form discussed in Sec. 2.2.1.2. log(VOT) was modeled because VOT can only be positive (for English voiceless stops). Models were fit in `R` using `lmer()` in the `lme4` package.

[2] For example, studies where VOT in different registers are compared have found that the difference in VOT between voiced and voiceless stops decreases in conversational speech; thus, VOT is a less reliable cue to stop voicing in natural speech than would be thought from laboratory experiments. It could turn out to be the case that some of the smaller VOT differences observed in lab speech, for example between stressed and unstressed syllables, simply wash out in conversational speech.

Table 6.1: Input variables for the VOT static models.

Predictor	*Type*
Speaking rate	continuous
Following vowel	factor (`high`, `non-high`)
Place of articulation	factor (`/p/`, `/t/`, `/k/`)
Stress	factor (`stressed`, `unstressed`)
Following segment type	factor (`vowel`, `sonorant`)
Frequency	continuous
Speaker gender	factor (`male`, `female`)

The models' predictors are based on seven input variables (Table 6.1). Speaking rate was log-transformed because it can only take on positive values.[3] Following common practice, frequency was log-transformed due to the highly-skewed frequency distribution of words, so that the predictor included in the model (log-transformed frequency) will have an approximately normal distribution.

Several steps were taken to minimize collinearity. The two continuous predictors were centered (following a log transformation). Each two-level factor was converted to a numerical predictor, then centered. Place of articulation was Helmert-coded, such that the two resulting predictors correspond to the difference between /p/ and the mean of /t/ and /k/ (predictor 1), and between /t/ and /k/ (predictor 2).

The models include a fixed-effect predictor for the main effect of each input variable, to model the effect of each input on VOT. They also include two fixed predictors for the interaction between place of articulation and the following vowel, to model lengthened VOT for /t/ before high vowels (an effect observed during annotation).

The models also include a by-speaker random slope for each fixed effect predictor, to account for variation in the effect of each predictor across different speakers. By-speaker

[3] Note that we are only including *overall* speaking rate as a predictor, without regard to between-subject differences in speaking rate. Another possibility is to include an additional predictor, the subject's mean speaking rate. In general, including both within-group and between-group predictors is recommended for variables for which different groups have different mean values (Snijders and Bosker, 2011: Sec. 4.6). In our case, adding a fixed-effect predictor for a speaker's mean rate did not change either model's qualitative results, and the predictor's fixed-effect coefficient was not significant.

and by-word random intercepts are also included, to allow for speaker-specific and word-specific adjustments in VOT. No correlations between random effect terms are included.

The resulting models predict a token's VOT as a function of terms for the speaker, the host word, and the surrounding spurt. The speaker's term is their own mean VOT, in part determined by their gender. The word's term comes from the word's own mean VOT, in part determined by the place of articulation of its initial stop consonant, which segments follow the initial stop, the word's frequency, and the word's stress pattern; the exact effect of these factors varies by speaker. The last term comes from the speaking rate of the surrounding spurt, and this effect also varies by speaker.

6.1.1 Diagnostics and goodness of fit

Normality of residuals: In an initial fit of each model the distribution of residuals was highly non-normal. Tokens with residuals more than 2.5 standard deviations from the mean were trimmed from each dataset, comprising 2.3% of the data for each dataset. On re-fitting, both models had residuals much closer to normality.

Multicollinearity among fixed-effect predictors: All correlations between fixed effects are small for both models ($|r| \leq .19$), suggesting that no two predictors are correlated. The condition number of the model matrix for the dataset corresponding to the 11 predictors included in each model was $\kappa = 6.0$, indicating very little collinearity in the full set of predictors.

Goodness of fit: As discussed in Sec. 2.2.4, there is no single good measure of model quality for linear mixed models. The option used here is R^2, defined in terms of model likelihood relative to a base model with only an intercept term (Eq. (2.4)). In our case, $R^2 = .491$ for the manual model and .451 for the automatic model.

Table 6.2: Summary of fixed effects in the model of manual (top) and automatic (bottom) VOT measurements: coefficient estimates $\hat{\beta}$, standard errors, corresponding t-values ($\hat{\beta}$/se($\hat{\beta}$)), and empirical MCMC p-values.

Predictor	$\hat{\beta}$	s.e.($\hat{\beta}$)	t	p_{MCMC}
Intercept	4.0650	0.0372	109.27	0.0001
Speaking rate	-0.0957	0.0090	-10.63	0.0001
Following high V	0.0270	0.0111	2.43	0.0126
POA (p vs. t/k)	0.1047	0.0117	8.95	0.0001
POA (t vs. k)	-0.1042	0.0238	-4.38	0.0001
Stressed	0.0718	0.0120	5.98	0.0001
Following C	0.0847	0.0102	8.30	0.0001
Frequency	-0.0537	0.0097	-5.54	0.0001
Male speaker	-0.0486	0.0371	-1.31	0.1528
Following high:POA (1)	0.0251	0.0074	3.39	0.0001
Following high:POA (2)	-0.0214	0.0117	-1.83	0.0340

Predictor	$\hat{\beta}$	s.e.($\hat{\beta}$)	t	p_{MCMC}
Intercept	4.1003	0.0365	112.34	0.0001
Speaking rate	-0.0851	0.0104	-8.18	0.0001
Following high V	0.0428	0.0122	3.51	0.0004
POA (p vs. t/k)	0.0872	0.0119	7.33	0.0001
POA (t vs. k)	-0.1062	0.0218	-4.87	0.0001
Stressed	0.0312	0.0142	2.20	0.0520
Following C	0.0796	0.0117	6.80	0.0001
Frequency	-0.0616	0.0114	-5.40	0.0001
Male speaker	-0.0780	0.0360	-2.17	0.0334
Following high:POA (1)	0.0145	0.0083	1.75	0.0234
Following high:POA (2)	-0.0236	0.0137	-1.72	0.0306

6.1.2 Results

Fixed effects Table 6.2 shows each model's estimate for each fixed-effect coefficient, with its standard error, t-statistic, and significance. Significances are the empirical p-values obtained by sampling from the posterior of the model parameters (see Sec. 2.2.3). Each coefficient can be interpreted as a predicted change in log(VOT); equivalently, the exponential of a coefficient gives a predicted percentage increase in VOT.

The intercept is of course highly significant in both models, and corresponds to a VOT of 58.3 msec ($= \exp(4.065)$) in the manual model and 60.3 msec in the automatic model.

Because all predictors have been centered, the intercept can be roughly thought of as an average VOT: the predicted VOT when all other predictors are held at their mean values.

There is a large and highly significant effect of speaking rate. An decrease of 2 standard deviations in speaking rate is predicted to increase VOT by 21.0% ($= \exp(2 \cdot .0957)$) for the manual model and by 18.5% for the automatic model. Both models predict a large and significant place of articulation effect. The first fixed-effect coefficient indicates that the VOT for /t/ and /k/ is predicted to be 36.9%/29.8% higher than for /p/ (manual/automatic model). The second coefficient indicates that VOT for /t/ is predicted to be 23.1%/23.6% higher than /k/ (manual/automatic model). Recall that the usual ordering is /p/≤/t/≤/k/, so predicting /t/ to have higher VOT than /k/ is unexpected. The place of articulation effect is the models' most surprising prediction, and is discussed in more detail below.

The effect of a following high vowel is significant in the manual model and highly significant in the automatic model, corresponding to increases of 6.3% and 10.3% in VOT relative to a following non-high vowel. There is a significant interaction of the height of the following vowel with POA in both models, which is perhaps best described by examining the model predictions. Fig. 6.1 shows the automatic model's predictions for each POA:Vowel height combination (across 100 posterior draws); the manual model's predictions are similar. It is visually clear that a following high vowel increases VOT markedly for /t/, and has little effect on VOT for /p/ and /k/. This confirms the suspicion, that VOT was higher for /t/ before high vowels, which motivated including a POA:Vowel height interaction in the models.

Stressed syllables are predicted to have 19.3%/7.9% longer VOTs in the manual/automatic models, relative to unstressed syllables. The effect is highly significant in the manual model, and nearly reaches significance in the automatic model ($p = 0.052$). Following segment type and word frequency have highly significant effects in both models. In the manual/automatic model, VOT is predicted to be 21.0%/19.6% longer in a stop preced-

Table 6.3: Effects of dropping different sets of random effect terms from the VOT model. Top group: all random slopes corresponding to each input variable. Bottom group: random intercepts for Speaker and Word.

Predictor	χ^2	df	Pr(>χ^2)
Speaking rate	14.0	1	< 0.0001
High vowel (+ POA interaction)	12.4	3	0.006
Place of articulation (+ High V interaction)	188.3	4	< 0.0001
High vowel:POA	5.2	2	0.073
Stress	26.6	1	< 0.0001
Following segment	1.0	1	0.329
Frequency	4.1	1	0.044
Speaker (intercept)	350.9	1	< 0.0001
Word (intercept)	203.7	1	< 0.0001

ing a sonorant (/l/, /r/, /w/, /j/) than in a stop preceding a vowel, and a decrease of 2 standard deviations in word frequency is predicted to increase VOT by 11.3%/12.9% .

Finally, an effect of gender in the same direction is predicted in both models; it is significant in the automatic model ($p = 0.033$), but does not reach significance ($p = 0.15$) in the manual model. Women are predicted to have higher VOT than men by 10.0%/16.6% in the manual/automatic model.

The estimated population-level effects for the two models are thus broadly similar. All variables have the same qualitative effect on VOT in both models. Most have similar significances as well, but two (stress and gender) do not. Because the models' predictions are so similar, from here on we only discuss the automatic model's results, for simplicity.

Random effects To get a sense of how important different random effect terms are to the model, we can compare models with and without particular terms (see Sec. 2.2.3). Table 6.3 shows the effect of dropping different sets of random effect terms from the model. Each row corresponds to a test of the hypothesis that one set of random effects has no effect. (For example, the first row summarizes a test of the hypothesis that the slope for speaking rate does not differ between subjects.)

The by-speaker intercept makes a highly significant contribution to the model (row 8 of Table 6.3), indicating that speakers differ in their baseline VOT values, after accounting for other factors. The predicted standard deviation of speakers' intercepts is $\hat{\sigma}_s = 0.155$, meaning that very high-VOT speakers are predicted to have VOT 85.9% higher than very low-VOT speakers (2 s.d. above/below the mean, respectively). The significant by-speaker intercept agrees with Allen et al. (2003), who found that American English speakers differed in their baseline VOT values for single-word productions, after controlling for speaking rate. Allen et al. do not report a model parameter describing how much speakers vary, but do report the adjusted VOT values predicted by their model. The standard deviation of the log of these values is 0.093, which can be compared to our model's $\hat{\sigma}_s = 0.155$. The 95% highest posterior density interval for $\hat{\sigma}_s$ is (0.103,0.197), which does not include 0.093. Thus, we can say with some confidence that speakers differ in VOT more in our dataset than in Allen et al.'s, perhaps because speakers' native dialects might differ in VOT values, British speakers might differ more than Americans, or speakers might differ more in conversational speech than in laboratory speech.

The by-speaker random slope for speaking rate also significantly contributes to the model (row 1 of Table 6.3), indicating that speakers differ in the effect of speaking rate (syllables/second) on VOT. The predicted standard deviation is $\hat{\sigma}_{\text{rate}} = 0.036$, and speakers in the dataset have predicted slopes between -0.132 and -0.019. Thus, for all speakers VOT decreases as rate increases. The significant random slope for rate agrees with Theodore et al. (2009), who also found that by-speaker differences in speaking rate (American English speakers, single word laboratory productions). Quantitative comparison of the spread of slopes predicted by our model with Theodore et al.'s results is not possible because they use a different measure of speaking rate (duration of the following vowel). However, they also find that all speakers' slopes have the same sign.

The random slopes involving place of articulation make a highly significant contribution to the model, indicating that speakers differ in the effect of POA (and its interaction with the following vowel's height) on VOT (row 3 of Table 6.3). The significance of the random slopes for POA contrasts with Theodore et al. (2009), who found that speakers did not differ in the POA effect—specifically, the difference between VOT for /p/ and /k/. The contrast is magnified by the fact that the random slopes involving POA make a highly significant contribution to our model compared to other random slopes, and are not even close to significance for Theodore et al.. A possible explanation for the discrepancy, discussed further below, is that dialects of English differ in the POA effect, so the poly-dialectal speakers in our sample would as well.

The by-word random intercept makes a highly significant contribution to the model (row 9 of Table 6.3), suggesting that words differ in their intrinsic VOT values after controlling for other factors. The predicted standard deviation of the words' intercepts is $\hat{\sigma}_w = 0.142$, meaning that a very high-VOT word is predicted to have VOT 76.4% higher than a very low-VOT word (2 s.d. above/below the mean).

The random slopes for following vowel height and its interaction with POA make a highly significant contribution to the model (row 2 of Table 6.3). The predicted slope for following vowel height has the same sign for all speakers, so differences between speakers do not change the direction of the effect following high vowel on VOT. The POA:Vowel height interaction also differs by speaker, but is not discussed further.

Speakers also differ in their slopes for frequency (row 7 of Table 6.3), though the effect is barely significant. Lower VOT for higher frequency words is predicted for all speakers; they differ only in the magnitude of the effect.

Finally, speakers differ significantly in their slopes for stress (row 5). The model predicts longer VOT in stressed than in unstressed syllables for 18 speakers, and the opposite pattern for three speakers. There are no significant differences between speakers in the effect of following segment type.

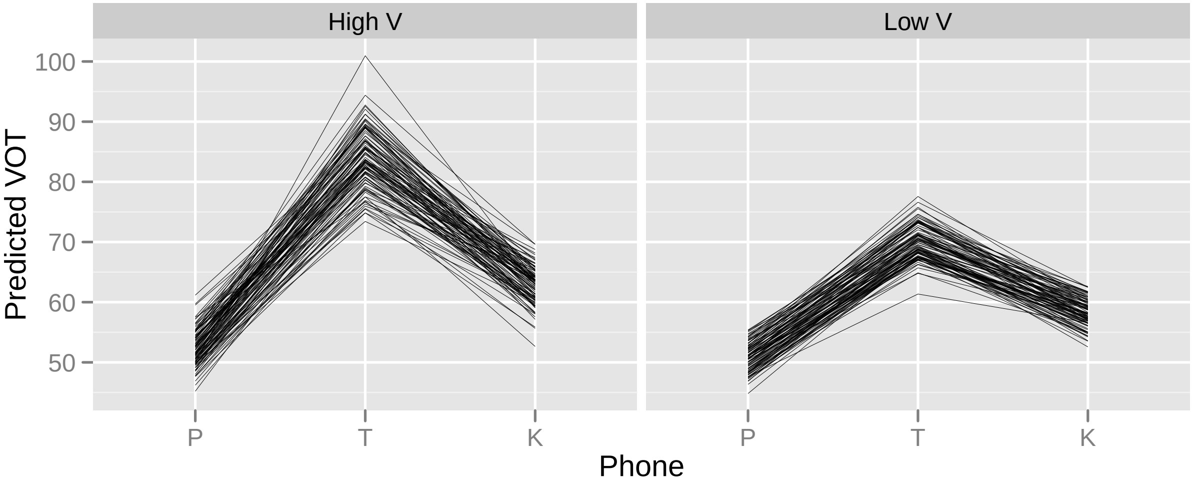

Figure 6.1: Predicted VOT by place of articulation and height of following vowel, for 100 draws of model parameters from their posterior distribution.

6.1.3 Discussion

Several aspects of the model are interesting with respect to previous work: the unexpected place of articulation effect, the relative effect of different predictors, and the total amount of VOT variation predicted by the model.

Place of articulation The most unexpected aspect of the model given previous work is the place of articulation effect. In many previous studies of English (primarily American English) and other languages, the effect of POA on VOT has been found to be consistent with the ordering /p/≤/t/≤/k/. In contrast, our model predicts /p/</k/</t/. Before discussing the meaning of this finding, we will check that it could not be due to uncertainty in the model's predictions, or to particular speakers with anomalous POA effects.

To quantify uncertainty in the model's predictions, we sample 100 times from the posterior distribution of the model parameters (see Sec. 2.2.3). Each draw gives a value for each model parameter. The values for the fixed effect coefficients can be used to obtain the predicted VOT values for /p/, /t/, and /k/ before a high or non-high following vowel (six numbers). The variability seen in these predictions indicates the uncertainty in the model's predicted population-level POA effect. Fig. 6.1 plots the predicted values for each draw. There is clearly significant variability in each of the six predictions, perhaps 10-20 ms for each. But there is little uncertainty about the *direction* of the POA effect: /p/</k/</t/ is predicted for each draw, regardless of the following vowel's height.

The same sampled model parameters can be used to examine uncertainty in the predictions of individual housemates, and to give a sense of whether a subset of housemates is responsible for the population-level POA effect. Fig. 6.2 shows the percentage of posterior draws where each possible POA order is predicted, for each housemate. In this figure, the closer to 100 a housemate's bar is for order x, the more confidently we can say that housemate's predicted effect "is x." For the majority of speakers, the predicted pattern is /p/</k/</t/. For some speakers (Belinda, Dennis, Jennifer, Rex), a non-trivial minority of draws give /p/</t/</k/, so we might say the predicted order is /p/</t/=/k/. Similarly, the predicted order for Michael is also /p/</t/=/k/ and the predicted order for Kathreya is /p/=/t/</k/. Thus, the majority of speakers show the same unexpected pattern (/p/</k/</t/) as the fixed effect, none show the expected pattern (/p/</t/</k/) and we can be confident that the model's population-level prediction is not due to a small number of speakers. Also, for most housemates there is relatively little uncertainty in the predicted POA effect, suggesting that the model's qualitative prediction for the effect of POA would not change with more data.

With some confidence in the /p/</k/</t/ effect, what should we make of it? Although this pattern is unexpected given previous work on VOT, it is not unattested: one language in Cho and Ladefoged's survey of VOT in 18 languages has markedly higher VOT for alveolar stops than for other places of articulation (bilabial, dental, velar).[4] In some studies of VOT in English, some individual speakers robustly show the /p/</k/</t/ ordering (e.g., Suomi, 1980: 69). And at least one study of VOT in British

[4] However, note that Cho and Ladefoged treat the alveolar stops in this language (Dahalo) as an anomaly because they are possibly affricated.

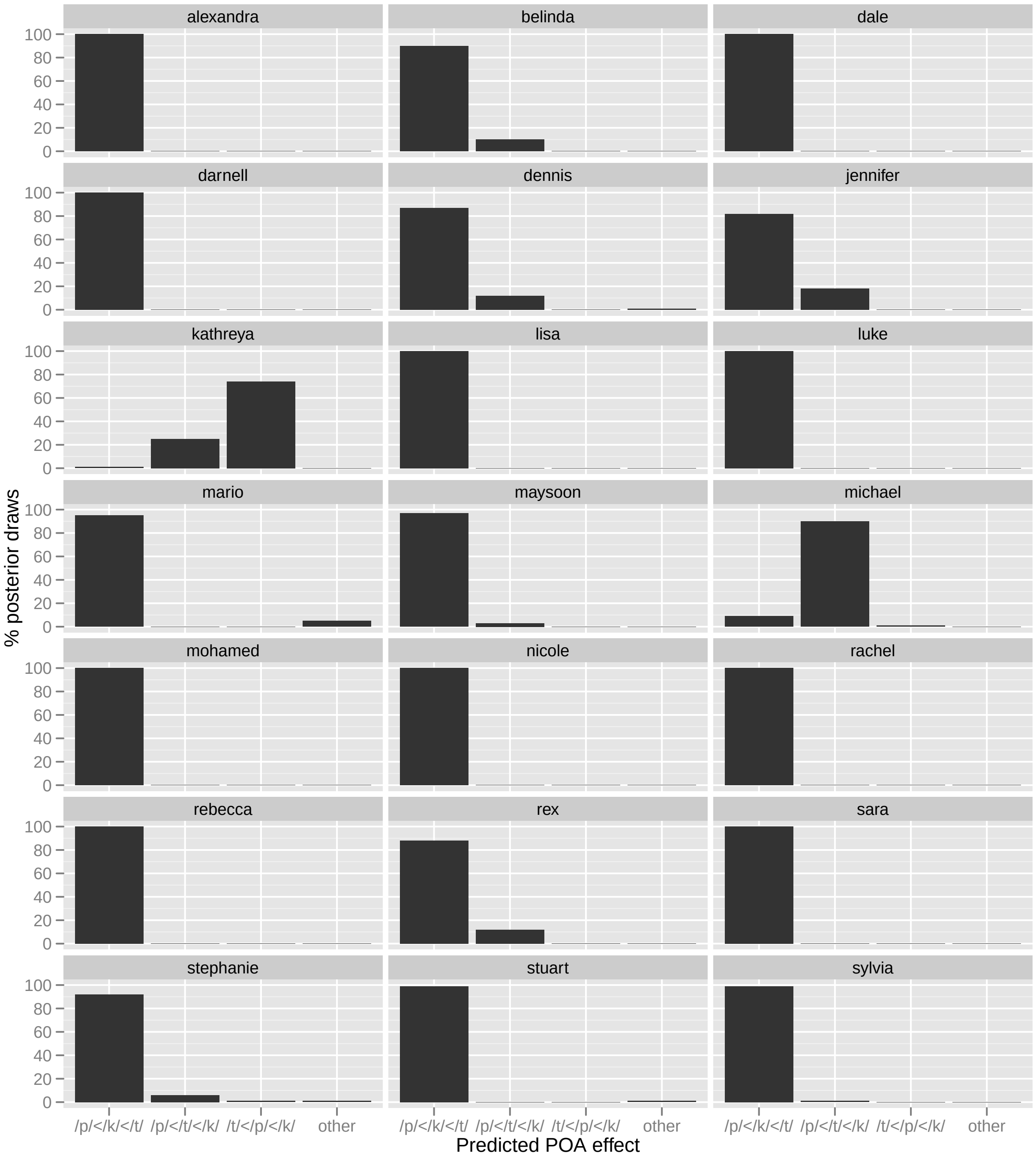

Figure 6.2: Percentage of draws with each ordering of predicted VOT by place of articulation, for 100 draws of model parameters from their posterior distribution. “Other” denotes /t/</k/</p/, /k/</p/</t/, or /k/</t/</p/.

dialects show a /p/</k/</t/ ordering in empirical mean values: in their study of adults in Sheffield (Northern England), Whiteside and Irving (1997) reported 37/69/56 msec for men and 53/78/66 msec for women. One possibility is that at least some British dialects have this marked ordering as part of their language-specific phonetics. In a similar vein, Docherty (1992: 139), in the most comprehensive study to date of VOT in standard British English (RP), suggests a language-specific VOT effect of /p/</t/=/k/. Because speakers in our dataset represent a variety of dialects, more work would be needed to determine if the POA effect observed here is a property of particular British dialects. Interestingly, recent work in British sociophonetics (discussed above) suggests that the VOT ranges used for different sets of stops varies greatly between British dialects, at least in Scotland and on the Scottish-English border, and may be related to social factors (Docherty et al., 2011; Scobbie, 2006; Watt and Yurkova, 2007). Our dataset may support the hypothesis that VOT is a particularly labile phonetic cue across the UK.

Comparing predictors Studying VOT in a corpus allows us to compare the effects of different factors on VOT, and assess how the model's predictions fit with the experimental literature.

Fixed effects: From the discussion above, we have some sense of the relative strength of different predictors, in terms of the percentage change in VOT that results from changing each one. Speaking rate and POA have the largest effects, followed by the following segment type and gender, followed by word frequency, the following vowel's height, and syllable stress. Speaking rate and POA are the factors that have been considered to have the largest effects in the experimental literature, so it is reassuring to see that they also have the largest effect when many other factors are considered. The weakness of the following vowel height effect may be related to the general uncertainty in the experimental literature on exactly which following vowels lead to higher VOTs. The small gender effect agrees with some previous work where women are found to have slightly higher VOT.

To our knowledge this is the first study to find an effect of gender on VOT when speaking rate is controlled for. However, we should perhaps not have much confidence in this effect, given that it was not significant in the manual model.

Random effects: From the discussion above of Table 6.3, we have a sense of the relative importance of different random effect terms to the model. First, consider the by-speaker and by-word random intercepts. That VOT varies between speakers replicates previous findings, while the finding that VOT varies between words is novel. The model predicts large differences between speakers and between words, meaning there is significant variation in VOT left to account for at the level of speakers and words. Further work is needed to determine whether this variation corresponds to random noise, or is somehow structured. The high by-speaker variation could mean speakers differ arbitrarily, opening the possibility (raised by Allen et al., 2003) that VOT could be used as an indexical feature by listeners to help perform speaker identification. On the other hand, it could be the case that by-speaker VOT differences are predictable, and we have not found the right predictors yet.[5] A similar picture holds for by-word variation: different words could have intrinsically different VOT values, a type of 'word-specific phonetics' (Pierrehumbert, 2002), or there may be important predictors accounting for at least some of the variance that we have not incorporated into the current model.

Next, consider the by-speaker random slopes. Assessed in terms of importance to model likelihood, by-speaker variation in the POA, stress, and speaking rate effects are the most important, followed by following vowel height and word frequency, while the following segment effect does not differ by speaker. However, the between-speaker differences are rarely large enough to predict that speakers differ qualitatively (different sign) in the effect of some factor. By and large, speakers differ *quantitatively* in the factors conditioning VOT variation, but not *qualitatively*.

Finally, it would be nice to know how much each variable contributes to the model,

[5] Of course, these possibilities are not mutually exclusive.

Table 6.4: Effect of dropping different sets of fixed and random effect terms on R^2, relative to R^2 for the full model. Top group: all terms corresponding to each input variable. Middle group: All random effects at the Speaker or Word level. Bottom: all fixed and random effect terms.

Predictor	Terms	$\Delta(R^2)$	$\%\Delta(R^2)$
Speaking rate		0.02764	6.12
High vowel		0.00318	0.70
POA		0.02060	4.56
High V:POA		0.00080	0.18
Stressed	Fixed effects, random slopes	0.00465	1.03
Following segment		0.00064	0.14
Frequency		0.00116	0.26
Gender		0.00000	0.00
All		0.07830	17.34
Speaker	Random intercept	0.02104	4.66
	Random intercept & slopes	0.16633	36.84
Word	Random intercept	0.06572	14.56
All	Fixed effects, random intercepts & slopes	0.45145	100.00

in terms of both the fixed and random effects. One way to quantify a variable's net effect is the difference in R^2 (denoted ΔR^2) resulting when all terms involving the variable are dropped from the model.[6] Table 6.4 shows the effect on R^2 of dropping different sets of fixed and random effect terms. By this measure, by-speaker and by-word variation make the biggest contributions to the model, followed by speaking rate and POA, then stress and following vowel height, then frequency and following segment type, and finally gender. Disregarding by-speaker and by-word variation, the ordering of different factors is roughly: speaking rate > phonological predictors > word-specific predictor (frequency) > social factor (gender), though there is admittedly only one variable for each of the last two groups.

[6] Note that ΔR^2 should be understood simply as a measure of change in model quality, not as a measure of change in the amount of variance explained (see Sec. 2.2.4).

Joint effect of predictors In addition to comparing the effects of different terms in the model, we can also consider how well they jointly account for VOT variation in the data. As noted above, the percentage change in data variance under the full model is $R^2 = 0.451$, meaning the model's likelihood improves over an intercept-only model by about 45%.

This figure can be compared to the value of $R^2 = 0.90$ found by Allen et al. (2003), in their laboratory study of isolated word productions, for a model containing a by-speaker random intercept and a fixed effect of speaking rate.[7] We can also compare less directly with Yao (2009b), who modeled VOT variation as a function of a variety of predictors (including speaking rate) for two speakers in the Buckeye Corpus. Yao fit a classic linear regression to each speaker's data, and found that $R^2 < 0.20$ for each speaker.

Despite the methodological difference between the three studies (ours, Yao, Allen et al.), it is striking that in Allen et al.'s study has the *highest* R^2 with the fewest variables, while in our study and Yao's, R^2 is much lower despite using more variables.[8] It seems reasonable to assume that this difference results from the different types of speech examined: conversational speech in our study and Yao's, vs. isolated words in a laboratory setting for Allen et al. Under this assumption, it seems that VOT in conversational speech is much *less predictable* than lab speech (c.f. Keating, 1998). : despite incorporating many known factors determining VOT variation, we explain only 1/2 of data variability in our conversational dataset; incorporating only a few factors, Allen et al. explain 90% of data variability in a laboratory dataset. There is also a huge difference in the effect of speaking rate in the two datasets: this variable mops up most of data variability for Allen et al., but only accounts for a modest amount of the variability observed in our dataset. This difference suggests a worrying possibility: some factors that influence VOT in conversational

[7] It is not clear exactly how Allen et al. calculate this figure, but the definition of R^2 we are using seems the most likely option.

[8] If Allen et al. had incorporated place of articulation into their model (they use words beginning with /p/, /t/, and /k/), their R^2 would be even higher and the difference to the other two studies would be more stark.

speech may not be detected in laboratory speech, since their effects may be very small compared to a few very large effects (speaking rate, speaker identity).

In general, our study and Yao's show that there is much VOT variation left to explain in conversational speech. Future work should see how much more variability can be accounted for by incorporating more predictors, or using different types of models (e.g., a model other than mixed-effects linear regression, or allowing non-linear effects of some predictors). It is an open question whether VOT—and phonetic variation more generally—will turn out to be largely predictable in conversational speech once the appropriate variables are plugged into the right model, or whether there is simply a lot of unpredictable noise.

6.2 Coronal stop deletion

Word-final t/d realization in the Big Brother data was modeled using mixed-effects logistic regression models, of the form discussed in Sec. 2.2.1.3, with crossed random effects for speakers and wordforms. Models were fit in R using the lme4 package (Bates et al., 2011), as linear mixed-effects models with a logistic link function. To maximize comparability of the results with previous work on CSD in British varieties, two models were fitted: one to data from all 17 speakers (the *full model/dataset*), the other to data from the 14 British speakers only (the *British model/dataset*). The models' predictors are based on seven input variables (Table 6.5).[9]

Following common practice, frequency (of the host word's CELEX wordform) was log-transformed due to the highly-skewed frequency distribution of words, so that the predictor included in the model (log-transformed frequency) will have an approximately normal distribution. Several steps were taken to minimize collinearity. Log frequency

[9] Note that there is only the possibility of the next word beginning with /t/ or /d/ if the following context is not a vowel. Thus, to avoid collinearity, the predictor for neutralizing context was set to 0 for tokens with a following vowel, and its actual interpretation is "next word begins with /t/ or /d/, given that the next word does not begin with a vowel."

Table 6.5: Input variables for the CSD static models.

Input	*Type*
Preceding context	factor (sibilant, sonorant, obstruent)
Following context	factor (consonant, vowel, pause)
Morphological context	factor (monomorpheme, semi-weak, weak)
Frequency	continuous (log-transformed)
Cluster voicing	factor (same, different)
Neutralizing context	factor (yes, no)
Speaker gender	factor (male, female)

was centered at the word level. Each two-level factor was converted to a sum-coded numerical predictor (-0.5, 0.5). Each three-level factor was Helmert-coded, such that the two resulting numerical predictors correspond to (for example) the difference between monomorphemes and the mean of semi-weak and weak pasts (predictor 1), and between semi-weak pasts and weak pasts (predictor 2).

In addition to the intercept, the models include fixed-effect predictors for the main effect of each input variable, to model the effect of each input on CSD rate. They also include a by-speaker random slope for each fixed-effect term, to account for variation between speakers in the effect of each predictor. By-speaker and by-word random intercepts are also included, to allow for speaker-specific and word-specific adjustments in CSD rate. No correlations between random effect terms are included.

The resulting models predict that a token's probability of deletion is a function of terms for the speaker, the host word, and the observation. The speaker's term is their own mean (log-odds of) deletion rate, in part determined by gender. The word's term comes from its own mean log-odds of deletion rate, in part determined by its preceding phonological context (including whether the final cluster has homogenous voicing), its CELEX frequency, and its morphological class; the exact effect of these factors varies by speaker. The observation-level term comes from the following phonological context (including whether it is neutralizing), and this effect also varies by speaker.

6.2.1 Diagnostics and goodness of fit

Normality of residuals: In an initial fit of each model using the full dataset, the distributions of residuals were highly non-normal, with much longer left and right tails than a normal distribution. Tokens with residuals more than 3 standard deviations from the mean were trimmed from each dataset, comprising 2.0% of the full dataset and 2.1% of the British dataset. On re-fitting, both models had residuals much closer to normality.

Multicollinearity among fixed-effect predictors: All correlations between fixed effects are small for both models ($|r| \leq 0.4$), suggesting that no two predictors are correlated. The condition numbers of the fixed effects was $\kappa = 7.6$ for both models, indicating very little collinearity.

Goodness of fit: A popular measure of overall model fit for logistic regressions is the Nagelkerke pseudo-R^2, written R^2_ν (see Sec. 2.2.4). The model for the full dataset has $R^2_\nu = 0.51$ relative to a baseline model consisting only of the intercept, and $R^2_\nu = 0.34$ relative to a baseline consisting only of by-speaker and by-word intercepts. The figures for the British dataset are $R^2_\nu = 0.56$ and $R^2_\nu = 0.38$. These values indicate that adding the models where all predictors are included substantially improve over baseline models where the same deletion rate is assumed across speakers and words, or where different speakers and different words have characteristic deletion rates.

6.2.2 Results

We discuss the models' predictors for each input variable, making reference to three tables. Table 6.6 shows the estimated fixed-effect coefficients for each model, with associated standard errors and significances. Table 6.7 shows how much each predictor contributes to each models, in terms of the effect on model likelihood of dropping both its corresponding fixed-effect and by-speaker random slope terms. Table 6.8 similarly shows how much each random effect term (by-speaker random slopes and intercepts)

Table 6.6: Predictors, standard errors, associated Wald's z-score, and significance for all fixed effect terms in the models for all speakers and for British speakers only

British speakers	$\hat{\beta}$	SE($\hat{\beta}$)	z	Pr(>|z|)
Intercept	-2.16	0.19	-11.2	<0.0001
Following (V vs. C)	1.11	0.11	10.1	<0.0001
Following (Pause vs. V/C)	1.05	0.07	15.1	<0.0001
Preceding (Sonorant vs. Sibilant)	0.27	0.15	1.8	0.069
Preceding (Obstruent vs. Son/Sib)	0.24	0.09	2.6	0.009
Frequency	0.22	0.09	2.4	0.018
Morph. class (Reg. vs. Irreg.)	-0.18	0.20	-0.9	0.38
Morph. class (Mono. vs. Past)	0.17	0.07	2.4	0.016
Voicing	0.49	0.43	1.1	0.26
Neutralizing context	0.56	0.06	9.2	<0.0001
Female speaker	0.27	0.29	0.96	0.34
All speakers	$\hat{\beta}$	SE($\hat{\beta}$)	z	Pr(>|z|)
Intercept	-1.74	0.24	-7.3	<0.0001
Following (V vs. C)	1.00	0.085	11.7	<0.0001
Following (Pause vs. V/C)	0.93	0.081	11.5	<0.0001
Preceding (Sonorant vs. Sibilant)	0.31	0.14	2.1	0.032
Preceding (Obstruent vs. Son/Sib)	0.25	0.073	3.4	0.001
Frequency	0.17	0.09	1.8	0.067
Morph. class (Reg. vs. Irreg.)	-0.02	0.17	-0.1	0.907
Morph. class (Mono. vs. Past)	0.19	0.076	2.5	0.012
Voicing	0.58	0.37	1.6	0.11
Neutralizing context	0.47	0.084	5.7	<0.0001
Female speaker	-0.08	0.43	-0.2	0.85

contributes to each model.

Intercept The intercept fixed effect term can be interpreted as the predicted deletion rate when all predictors are set to zero. Because all predictors have been centered, the intercept can be roughly thought of as an "average rate" of CSD, after controlling for effects of linguistic and social conditioning factors, and for adjustments due to individual speakers and words. Transforming from log-odds, the average rates are 10.3% (=logit^{-1}(-2.16)) for British speakers and 14.9% for all speakers. These percentages are significantly lower the mean of the deletion rates for each (speaker,word) pair, which are 42.1% (British) and

Table 6.7: Effect of dropping each predictor (fixed effects + random slopes) from the model, for all speakers (left) and for British speakers only (right).

Predictor	df	British		All	
		χ^2	Pr(>χ^2)	χ^2	Pr(>χ^2)
Following context	4	1067.9	<0.0001	1174.3	<0.0001
Preceding context	4	30.5	<0.0001	60.2	<0.0001
Morphological class	4	5.6	0.228	10.2	0.033
Frequency	2	10.1	0.007	17.2	<0.0001
Voicing	2	90.9	<0.0001	108.1	<0.0001
Neutralizing context	2	104.7	<0.0001	104.5	<0.0001
Gender	1	0.8	0.36	0.0	0.85

Table 6.8: Effect of dropping random slope (top) and random intercept (bottom) terms from the models for all speakers and for British speakers only.

Predictor	df	British		All	
		χ^2	Pr(>χ^2)	χ^2	Pr(>χ^2)
Following context	2	23.4	<0.0001	35.1	<0.0001
Preceding context	2	17.4	<0.0001	37.3	<0.0001
Morphological class	2	0.2	0.915	1.2	0.541
Frequency	1	2.8	0.095	11.3	<0.0001
Voicing	1	76.7	<0.0001	94.5	<0.0001
Neutralizing context	1	0.0	1	11.1	0.001
Speaker	1	11.0	0.001	51.5	<0.0001
Word	1	229.5	<0.0001	264.0	<0.0001

44.6% (full). It is difficult to compare these mean deletion rate numbers to those reported in previous studies of CSD, especially because of the non-standard choices taken here on which tokens to discard.

The standard deviations corresponding to the by-speaker and by-word random intercepts are $\sigma_s = 0.44$ and $\sigma_w = 0.91$ for the British data, and $\sigma_s = 0.79$ and $\sigma_w = 0.86$ for the full dataset. Unsurprisingly, the full set of speakers differs more in CSD rate than British speakers alone do. Words have about the same spread of deletion rates in the British dataset as in the full dataset. For both datasets the amount of between-speaker variation in deletion rate is smaller than the amount of between-word variation; for the British data, much smaller.

The relative size of σ_s^2 and σ_w^2 can be thought of as measuring how much the remaining variation in CSD usage can be attributed to the speaker vs. the host word, after taking factors into account that were included in the model as predictors. By this measure, speakers account for 1/4 as much variation as words (σ_s^2/σ_w^2). This discrepancy is striking given that *no* speaker-level predictors except gender (which is not significant for either dataset) have been included in the model; most predictors are word-level. What is not clear is to what extent the remaining by-speaker and by-word variances indicate truly arbitrary differences in CSD rate between speakers and words, or the effects of factors not modeled here. The size of σ_w relative to σ_s suggests that searching for word-level predictors will be a more fruitful direction for future work modeling CSD in Big Brother.

To my knowledge, Johnson's (2012) study of the Buckeye Corpus is the only CSD study where by-speaker and by-word random intercept variances are reported. He also finds less by-speaker variation than by-word variation: $\sigma_s = 0.48$, $\sigma_w = 0.59$.[10]

[10] Recall that the Buckeye Corpus consists of sociolinguistic interviews with speakers from a single community (Columbus, Ohio). Note that Johnson defines "word" orthographically, while here it is defined as the CELEX wordform: the orthographic word augmented with non-orthographic (e.g., morphological, part-of-speech) information. However, re-fitting the Big Brother models using the orthographic definition of word gives qualitatively similar results.

Phonological context The fixed-effect coefficients for preceding and following phonological contexts, as well as their significances, have similar values in the two models. The following context coefficients are highly significant, while one preceding context coefficient is somewhat significant (p=0.03–0.06) and the other is very significant (Table 6.6). As expected from previous work, deletion is more likely before consonants than before vowels, as well as progressively more likely following sibilants, sonorants, and obstruents. Deletion is least likely before pauses, as found in some previous studies. Comparing to British studies in particular, TT find deletion least likely before pauses, while Smith et al. find it least likely before vowels.

The random effect coefficients for both preceding and following phonological contexts also make highly significant contributions to each model (Table 6.8), meaning that speakers differ in their expected deletion rates in different phonological contexts.

Morphological class The estimated fixed-effect coefficients for morphological class are similar for the two datasets. There is no significant difference between the deletion rate of irregular and regular past tense forms. However, there is a significant difference between monomorphemes and past tense forms, in the expected direction (monomorphemes delete more than pasts). For both models, the contribution of the random slopes for morphological class to the model is not significant. That is, although speakers show a morphological effect as a group, they differ very little in the size of the morphological effect.

Frequency For both datasets, there is an effect of frequency in the expected direction: more frequent words undergo CSD more often. The effect is significant ($p = 0.018$) in the British model and marginal ($p = 0.067$) in the full model. In addition, the random slope for frequency makes a significant contribution to the full model (<0.0001), but not the British model ($p = 0.095$). These differences between the models are due to a single speaker (Mohamed) who has a far lower slope for frequency than other speakers. A model fit to the full dataset with his data removed gives a much more similar frequency effect to the British model. For both datasets, the predicted frequency effects for each speaker differ in magnitude, but are all in the expected direction.

Voicing For both datasets, the fixed effect coefficient for voicing of the final cluster is in the expected direction (higher deletion rates for hetero-voiced clusters), but does not reach significance. Note that the actual estimated effect of voicing is actual fairly large, compared to the coefficients for other fixed effects: for the British data, for example, having a hetero-voiced cluster makes the odds of deletion 1.6 times higher (=exp(0.49)). The effect does not reach significance because the corresponding standard error is large, which is due to huge differences between speakers in the effect of voicing. For both datasets, the random slope for voicing makes a far more important contribution to the model than the random slope for any other predictor, assessed by the effect of dropping it from the model. Importantly, it is not just the magnitude of the effect that differs by speaker, but its direction: some speakers are predicted to delete more for homo-voiced clusters, and others less.

Neutralizing context There is a highly significant effect of neutralizing context in the expected direction in both models: the odds of deletion are 1.6/1.75 higher (full/British data) before a word beginning with a coronal stop. British speakers differ little in this effect, with a random slope of near zero, while the full set of speakers differs somewhat (but with no speaker having a predicted effect in the unexpected direction). Accordingly, the contribution of the random slope term is not significant for the British model, but highly significant for the full model.

Gender Neither model has a significant gender effect: there is no evidence that women delete at a different rate than men.

6.2.3 Discussion

Several aspects of the models are interesting in light of previous work. In addition, there are several methodological points which may be important for explaining discrepancies between our models and previous work, and for future work on CSD.

Caveat It should be noted from the outset that our datasets differ from those used in most sociolinguistic studies in a crucial respect: they do not consist of speakers from a single speech community, but a collection of speakers from very different dialect regions. Because dialects differ to some extent in the effects of different conditioning factors on CSD (Schreier, 2005)—in particular the relative strength of different factors—modeling data from geographically dispersed speakers may lead to somewhat different results than modeling data from a single speech community. To maximize comparability with previous work, we only consider the British model in our discussion.[11]

Morphological class Morphological class was coded as two fixed effects in our model: one for the difference between monomorphemes and pasts, and one for the difference between weak and semi-weak pasts, which we write as $\beta_{M/P}$ and $\beta_{SW/W}$ for convenience. Monomorphemes were found to be deleted significantly more often than pasts ($\beta_{M/P}$: p=0.01), in agreement with most previous work on CSD. However, there was no significant difference in deletion rate between irregular and regular pasts ($\beta_{SW/W}$: $p = 0.38$). Importantly, the difference in significance between the two coefficients is due to the standard error of $\beta_{SW/W}$ being much higher, despite the coefficient estimates being of similar magnitude. That is, the size of both effects is comparable, but the model is "less certain" about the semi-weak/weak effect, due to the sparsity of semi-weak pasts in the data (110 tokens, 41 wordforms), relative to both weak pasts (338 tokens, 198 wordforms) and monomorphemes (4266 tokens, 292 wordforms). This imbalance is due to the

[11] Data spanning a single country is in fact not unprecedented in the CSD literature: for example, Guy (1991) and Neu (1980) pool speakers from across the US.

frequency with which irregular pasts occur in English, and thus can be assumed to be present in CSD datasets used in previous work, which are typically of comparable size to our dataset (thousands of observations). Almost all previous work (where multifactorial analyses are reported) reports results in the form output by VARBRUL/GOLDVARB, which does not include standard errors on coefficient estimates. Rather, if the model selects a factor as significant (using a stepwise model selection procedure), estimated "factor weights" are output for all levels of the factor, no matter how certain the model is about the factor weight for each level. Our results suggest a possible source of at least some of the divergent results about CSD in semi-weak pasts reported in the literature, compared to the consistency of results about monomorphemes vs. pasts: the factor weights reported for semi-weak pasts are simply less reliable, due to data sparsity.

Although we do find a morphological effect in the data, it is very weak. Dropping the morphological class terms (2 fixed effects, 2 random slopes) from the British model does not make a significant difference in model likelihood, relative to the degrees of freedom (Table 6.7). To compare more directly with previous work, where random slopes have not been included, we can consider a model where the two random slope terms have been removed. Dropping the two fixed effect terms from this model only has a marginal effect ($\chi^2(2) = 5.5$, $p = 0.065$), suggesting that morphological class is at best a minor factor in explaining the CSD data.[12] How can this be squared with the strong morphological class effects found in most previous studies?

One possibility, following TT's finding of no morphological effect on CSD for their York speakers, is that morphology has a much weaker role (if any) in British varieties, compared to North American varieties. Another possibility involves the relationship between morphological class and frequency. On average in English, monomorphemes have higher frequency than pasts, and irregular pasts have higher frequency than regular

[12] Note in particular that in a stepwise procedure with $p = 0.05$, the cutoff most common used in VARBRUL/GOLDVARB analyses, Morphological Class would not be selected as significant.

pasts (as noted in the context of CSD by Myers and Guy, 1997: 220). Thus, words from morphological classes that promote deletion tend to have higher frequencies, which also promotes deletion. A model including both variables estimates how important each input is (frequency, morphological class) in the presence of the other. In our case, frequency makes a more significant contribution to the model, as measured by the effect of dropping fixed+random effect terms (Table 6.7). If frequency is not included in the model, we are in danger of *omitted variable bias*: when two variables are positively correlated, omitting one from the model can spuriously increase the significance of the other. This is the case for the British data: if frequency (both fixed and random effect terms) is left out of the model, the morphological effect becomes highly significant ($\chi^2(5) = 578.6$, $p < 0.0001$). Many previous studies of CSD have not included frequency as a predictor. If the datasets used in previous studies showed an effect similar to the current dataset—where frequency has a significant effect, after controlling for morphological class—then the effect of morphological class may have been overestimated, due to not including frequency in the model.

Frequency In the British data, there is a significant effect of frequency on deletion rate. Compared to other predictors in terms of the significance of its contribution (fixed and random effect terms) to model likelihood, the frequency effect is more important than morphological class (which is not significant), and less important than all predictors related to the phonological context. This finding is similar to Johnson (2012), who finds that frequency mattered less (in terms of model likelihood) than phonological context and morphological class.

This finding contributes to a larger debate about the existence and size of frequency effects on sociolinguistic variables. Some studies have found frequency to be a significant predictor of CSD, and argue for the importance of frequency (Bybee, 2000, 2007; Jurafsky et al., 2001; Myers and Guy, 1997) As pointed out by Walker (2012), an issue with many of these studies is that the effect of frequency is considered without fully controlling for other variables known to affect deletion rate using a multivariate analysis, so it is not clear how much of a role frequency *per se* plays. Put otherwise, frequency effects may be overestimated (e.g., due to omitted variable bias, discussed above). Studies where frequency is considered in a multivariate analysis have found modest or negligible effects of frequency on sociolinguistic variables, including CSD (Abramowicz, 2007; Clark and Trousdale, 2009; Dinkin, 2008; Labov, 2010: Ch. 13; Walker, 2012). An issue with these studies, which mostly use VARBRUL/GOLDVARB (with the exception of Johnson, 2012), is that because these programs can only model discrete predictors, frequency is discretized into categories (e.g., low/high). Discretizing continuous predictors can lead to biased estimates of their effect on the response, and increase Type II error (the probability of failing to reject the null hypothesis of no effect, when in fact it is false) (MacCallum et al., 2002). Thus, discretizing frequency could lead to underestimation of its effect.

In sum, the effect of frequency may have been overestimated in (many) studies arguing for frequency effects, and underestimated in studies arguing against frequency effects, for methodological reasons. Given that the importance of frequency as a predictor of sociolinguistic variation relative to other predictors is of theoretical interest, correcting these methodological issues in future work is important to assess the true contribution of frequency.

Voicing There is no significant effect of cluster voicing on deletion rate in the British data. This result contrasts with the general finding that words with homo-voiced clusters have higher deletion rates. The explanation for this discrepancy may lie in our inclusion of a random slope term for cluster voicing. As discussed above, the model predicts huge between-speaker variance in the effect of voicing. If the random slope term is left out of the model, the fixed effect of voicing becomes significant. This suggests that words with homo-voiced clusters are being deleted more often, on average; but there are huge differences between speakers in the size and direction of the voicing effect, and once these are accounted for, the spread of speakers' deletion rates is too large relative to their mean value to conclude that there is a non-zero population-level effect. Thus, including a random slope term for cluster voicing makes a qualitative difference in what conclusions can be drawn from the model.

The voicing effect in our data can be interpreted with respect to previous work in two ways, depending on the (unknown) reason we observe such large between-speaker differences. The spread of voicing effects could be the result of speakers' dialects having different voicing effects, some even with the reverse effect (homo-voiced deleting less than hetero-voiced).[13] In this case, we would expect future work on British speech communities to find very different voicing effects. Cluster voicing would then join preceding phonological context as a factor in CSD that differs greatly across dialects. Another possibility is that large by-speaker variation in the voicing effect exists in every community. In this case, because previous studies that have considered voicing as a factor in CSD may have overstated its effect by not allowing for between-speaker variation. The second possibility, though remote, argues for incorporating between-speaker variation (i.e., via random slopes) into future studies of CSD.

6.3 Vowel formants

For each vowel, we build one linear mixed-effects model for normalized F1, and one model for normalized F2, resulting in six models in total.

Of course, F1 and F2 do not vary independently, and we would ideally model them *jointly*, for each vowel. It turns out that mixed-effects models for multivariate responses are highly non-trivial, and are the subject of ongoing research. Researchers have dealt with the problem of modeling F1/F2 data in a variety of ways. F1 and F2 can be jointly

[13] No previous work on CSD in British dialects has examined cluster voicing, so it is not clear how plausible this hypothesis is; however, the reverse effect has been found at least once (by Hazen, 2011, in Appalachia).

modeled in a non-mixed regression, and the importance of each predictor assessed by performing a multivariate analysis of variance (MANOVA) (Hay et al., 2006). In this approach, no random effect terms can be included. A particular transformation of F1 and F2 can be chosen based on the question of interest; for example, because GOOSE-fronting mostly involves change in F2, F2 alone can be modeled. Finally, separate regressions can be performed for F1 and F2. We choose the last option because we want a model that includes random effects, and because it is important for F1 and F2 to be able to vary arbitrarily (in particular in the dynamic models considered in the next chapter, so that longitudinal variation in any direction in F1/F2 space is possible). However, building models that respect both the multivariate and hierarchical nature of vowel formant data is a worthy direction for future work.

It is also worth noting that the static models for vowel formants are much more basic than those described for VOT and coronal stop deletion, both in terms of the number of conditioning factors considered and the amount of comparison with previous work. The vowel formant data turns out to be much sparser (fewer tokens, relative to the number of conditioning factors), forcing us to adopt simpler models. There is also less relevant previous work to which we can directly compare our results. Our goals for the static models are simply to build basic models that account for the effects of speaker identity and the consonantal context on vowel formants, and to check as much as possible that the models' results make sense with respect to previous work.

Predictors The terms included in our models are based on the coding given by FAVE, which parametrizes the surrounding segments as four input variables:

- *Preceding segment*: Oral apical, nasal apical, oral labial, nasal labial, liquid, obstruent+liquid cluster, palatal, velar, w/y, other.
- *Following segment manner of articulation*: Affricate, fricative, lateral, nasal, rhotic, stop, other.

- *Following segment place of articulation*: Alveolar, bilabial, interdental, labiodental, palatal, velar, other.

- *Following segment voicing*: Voiced, unvoiced, other.

This parametrization is taken from Plotnik, a program widely used by sociolinguists for visualizing the effects of context on vowel quality.[14] The coding scheme reflects most groups of segments which have been found empirically to condition variation and change in sociolinguistic studies (usually of American English, but also other varieties and other languages). We use this parametrization for convenience, without further comment.

Model specification Each of the six models was fit in `R`, using `lmer()` in the `lme4` package. Following consonant's voicing was Helmert-coded to give two fixed-effect predictors: one for the difference between unvoiced and voiced, and one for the difference between unvoiced/voiced and 'other' (for tokens in word-final position, or followed by another vowel). Each of the other three consonantal context variables was included as a random intercept term, as was the speaker's identity.[15] Thus, the model included four random intercept terms: by-following manner, by-following place, by-preceding segment, and by-speaker. This model structure means that we are assuming (for example) that for F2 of GOOSE, different preceding segment classes affect F2 to varying degrees, the effects of different preceding segment classes are normally distributed, and effects of preceding segment class are independent of effects of following manner and following place.

We note that in contrast to the models for VOT and coronal stop deletion, we have not included by-speaker random slopes or by-word random intercepts, because the data

[14] `http://www.ling.upenn.edu/~wlabov/Plotnik.html`

[15] The choice to code these consonantal context variables as random effects is motivated by their extreme non-independence. That they are not independent (for example, there is no interdental nasal in English) means that including them all as fixed-effect predictors would give a singular model matrix. A new, non-singular parametrization of the variables would need to be chosen, and it is not clear how to make this choice in a principled way.

Table 6.9: Fixed effect estimates, t-values, and associated significances for the effect of following consonant voicing on normalized formants of GOOSE, STRUT, TRAP′.

Vowel	F1			F2		
	$\hat{\beta}$	t	P(> $\|t\|$)	$\hat{\beta}$	t	P(> $\|t\|$)
GOOSE	-0.13	-2.2	0.029	-0.02	-0.16	0.87
STRUT	-0.23	-6.4	<0.0001	0.17	6.4	<0.0001
TRAP′	-0.08	-2.1	0.036	0.04	1.7	0.085

is simply too sparse. The models thus assume that the effects of consonantal context are the same for different speakers, and that words do not differ in their formant values, after accounting for by-speaker differences and the consonantal context. The latter assumption has the practical effect that higher-frequency words will have more influence on the model's estimates. However, it will turn out to be possible to include by-word random intercepts in the dynamic models (Chapter 7), where we fit models only for individual speakers.

Diagnostics and goodness of fit In an initial fit of each of the six models, the distribution of residuals was somewhat non-normal. We manually trimmed outliers (using histograms of residuals), comprising between 0.3% and 1.9% of the data for the six datasets. On re-fitting, all models had residuals much closer to normality.

Because each model contains only predictors for a single Helmert-coded input (following consonant voicing), there is no risk of collinearity.

Goodness of fit is assessed by R^2, defined in terms of model likelihood relative to a base model with only an intercept term (Eq. (2.4)). R^2 is 0.16 and 0.46 for F1 and F2 of GOOSE, 0.50 and 0.44 for STRUT, and 0.34 and 0.63 for TRAP′.

6.3.1 Results

Voicing Table 6.9 shows the estimated fixed effect coefficient for each model, along with its associated t value and significance, for the difference between a following voiced and following voiceless consonant. Recall that based on laboratory studies, we expect F1 to be lower for a following voiced consonant, and for following consonant voicing to have little effect on F1 (see Sec. 4.4.3.2). The prediction for F1 is met for all three vowels, with F1 predicted to be significantly lower when followed by a voiced consonant. The marginal and non-significant following voicing effects in the F2 models for GOOSE and TRAP′ are consistent with the prediction for F2; however, there is a large and unexpected effect of following voicing for STRUT.

Random effects Each model contains four random intercepts, corresponding to speaker, preceding segment identity, following segment place, and following segment manner. The variance of the by-speaker random intercept is large in each model, as expected given the range of qualities of the three vowels across speakers' dialect regions (see Sec. 5.2.4.2).

For the three consonantal context random intercepts, the best linear unbiased predictors (BLUPs; see Sec. 2.2.2) are plotted in Figs. 6.3–6.4, along with 95% confidence intervals from simulations over the posterior of the model parameters. For most terms (e.g., for following segment place for F2 for TRAP′), there is significant uncertainty about the relative position of the different random effects. This uncertainty reflects the small number of tokens per random effect in each dataset.

In general, it is difficult to compare the effects of consonantal context predicted by our models to previous work, because laboratory studies have only examined a subset of consonantal contexts (e.g., stops and fricatives, stops alone; see Sec. 4.4.3). For example, our `following velar` means "any following consonant with a velar place of articulation"; in laboratory studies, it has meant "non-nasal velar consonants" or "velar stops."

That said, a few comparisons with previous work are possible, and offer some reassurance that our models make sensible predictions. For GOOSE, the expected pattern is for F2 to be highest before coronals and /j/, and smallest before /l/ (Sec. 4.4.3.2). This is exactly the pattern predicted by our model of F2 for GOOSE (Fig. 6.3): for preceding con-

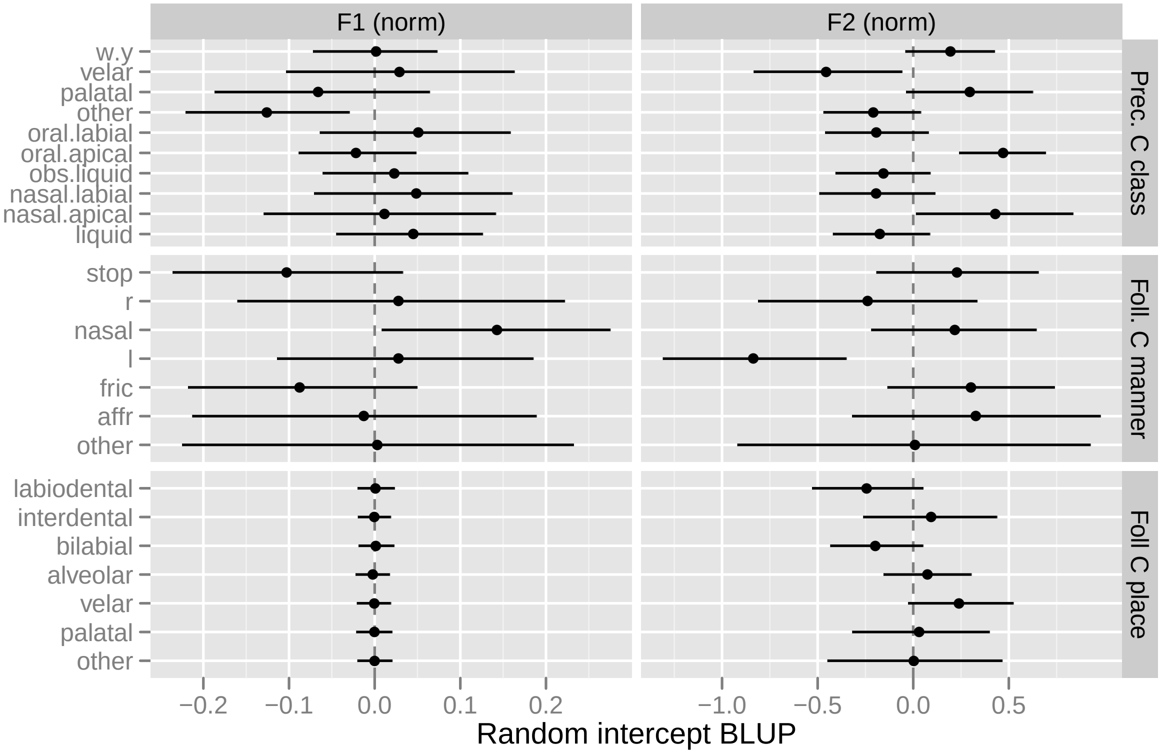

Figure 6.3: BLUPs for random effects corresponding to the preceding segment, following segment place, and following segment manner, in the models for normalized F1 and F2 for GOOSE. Errorbars indicate 95% HPD intervals based on 1000 draws from the posterior distribution over model parameters.

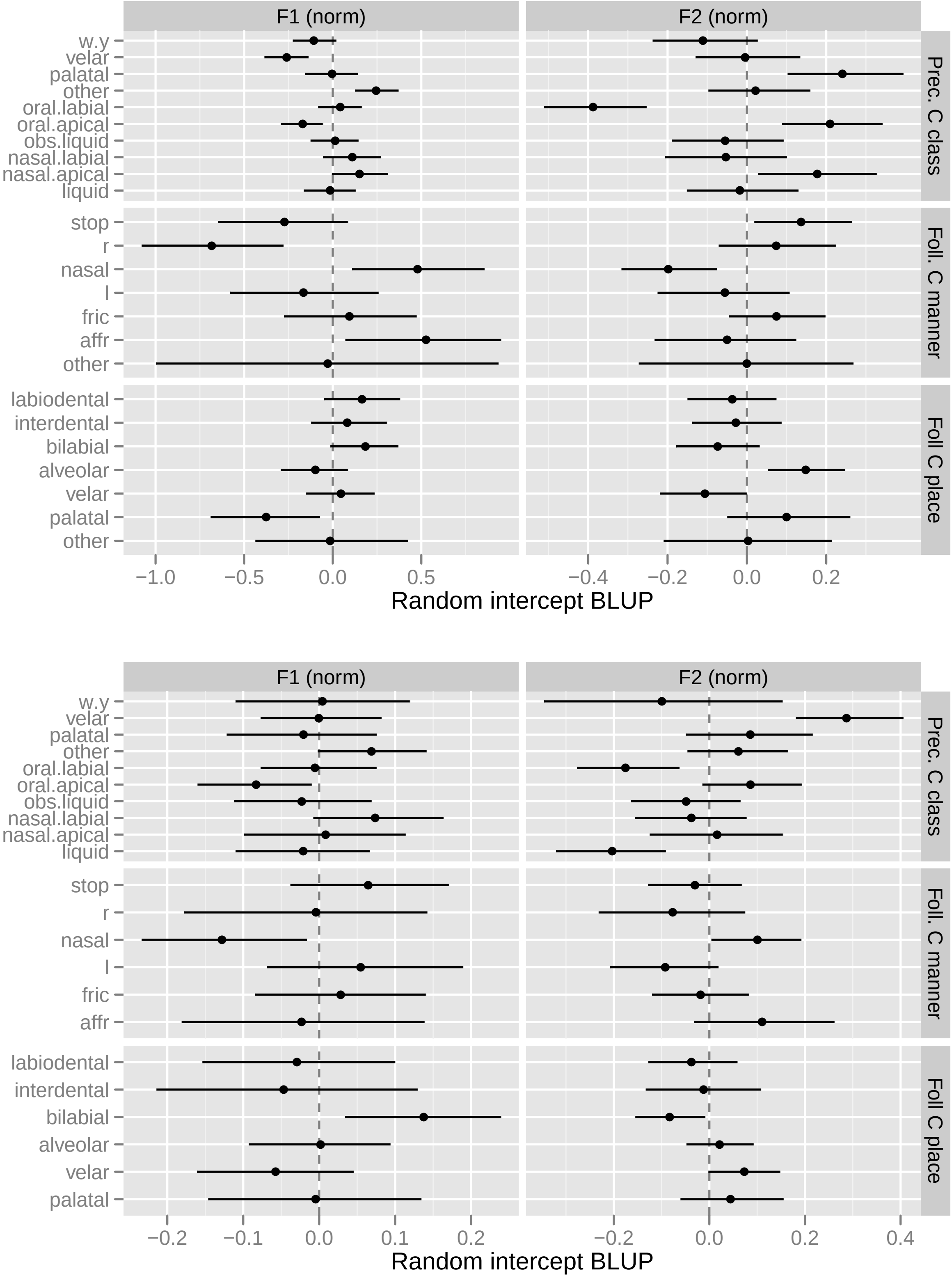

Figure 6.4: BLUPs for random effects for STRUT (top) and TRAP′ (bottom) F1 and F2 models. See Fig. 6.3 caption.

sonant class, the largest BLUPs are for `palatal`, `oral apical`, and `nasal apical`, which together make up the coronals; the next highest BLUP, is for `w/y`, which includes /j/. And for following consonant manner, the BLUP for /l/ has the lowest value.

We can also consider the effects of a following nasal on F1, the formant most affected by nasalization. For following consonant manner, the BLUP for `nasal` is positive for GOOSE and STRUT, and negative for TRAP′, compared to the BLUPs for other manners of articulation. This is the expected effect of coarticulation with a nasal: lower F1 for low vowels, and higher F1 for high vowels (Beddor, 1982).

CHAPTER 7

BIG BROTHER: DYNAMIC MODELS

Our goal in building dynamic models is to understand the dynamics of each variable within individuals, after controlling for linguistic factors. Ideally, we would treat time like any other variable, analogously to how we modeled the effects of linguistic factors: terms for group-level time trends (fixed effects) and individual deviations from the group-level trends (random slopes). Analogously to including a word-level random effect to model word-level linguistic predictors that are not included in the model, we would also allow for clip-by-clip variation (random intercept), to account for clip-level factors that are not included in the model (e.g., conversational topic). We would then be able to fit a single model for each variable, modeling linguistic factors and time across all speakers. However, modeling time like linguistic factors assumes that speakers show qualitatively similar time dependence, namely time trends of a particular functional form, and similar amounts of by-clip variation. We shall see that this is not the case for either type of time dependence: speakers differ qualitatively both in time trends and by-clip variation. Building a model where both are allowed to vary turns out to be highly non-trivial, particularly given the exploratory nature of this study.

Instead, we will fit models for individual speakers. The basic strategy will be to fit a series of models of a speaker's use of a variable that assume different types of time dependence, then pick the best one.

In this chapter, we first describe each step of this process: what types of time dependence are considered, how linguistic factors are controlled for, and how the best model is chosen (Sec. 7.1). We then describe the individual models selected for each variable (VOT, CSD, vowel formants), and examine whether each shows evidence of convergence (Sec. 7.2–7.4). Finally, we consider explanations for the varied patterns of time dependence that are observed (Sec. 7.5), and discuss implications of our results for the relationship between short-term and long-term change (Sec. 7.6).

7.1 Procedure

Types of time dependence We consider two basic types of longitudinal variation: random *by-clip variation*, without regard to when the clips occur, and *time trends*, corresponding to linear or non-linear trends in a variable's use as a function of what day the clip is from.[1] In modeling terms, by-clip variation corresponds to including a by-clip random intercept, and time trends correspond to fixed-effect terms for functions of EPISODE.

Because there is no previous work on medium-term time trajectories of linguistic variables, we have no a priori hypothesis on what the dynamics of our variables will be, and we must choose a set of possibilities to consider. We consider four types of time dependence, corresponding to the presence or absence of by-clip variation and time trends:[2]

- *Type 1*: No time dependence.
- *Type 2*: By-clip variation. (Model includes by-clip random intercept.)
- *Type 3*: Time trend. (Model includes fixed effect term(s) for a linear, quadratic, cubic, or quartic function of EPISODE.)
- *Type 4*: By-clip variation and time trend. (Model includes both types of term.)

Thus, for each variable for each speaker, we try 12 models (1 for Types 1–2 and 10 for types 3–4). The cutoff of degree=4 for polynomial time trends is empirically motivated: allowing higher-order terms tended to lead to the best model (according to AIC; see below) having an overfitted time trend with a high degree.

[1] Other types of time dependence are possible. For example, we are not accounting for the relative position of data points in a clip, or the fact that there are occasionally multiple clips from the same speaker in a single day.

[2] Note that this parametrization means we cannot fit *arbitrary* patterns of time dependence, such as non-independence of clips which occur on adjacent days, or periodic time trends. Non-trivial covariance structures for by-clip random effects and non-parametric time trends will be considered in future work, for example by using generalized additive models. Ultimately, the types of time dependence considered should depend on what longitudinal dynamics are observed in empirical studies of trajectories of linguistic variables, once more are conducted.

Clips occur on average between once per 3.1 days (median 2.6, range 1.8–7.7), depending on the speaker (Table 5.1), with greater density for speakers who are on the show for longer. (As housemates are evicted, the number of housemates whose clips can be aired decreases.) Thus, to distinguish between by-clip variation and long-term trends, we can only consider housemates who spend a relatively long period in the house. Fifty days (of 93 total) was chosen as an arbitrary cutoff, leaving 12 speakers (of 21): Dale, Darnell, Kathreya, Lisa, Luke, Michael, Mohamed, Rachel, Rebecca, Rex, Sara, and Stuart. These housemates account for 85% of housemate speech in all diary room clips.

Controlling for linguistic factors Linguistic factors are controlled for in the dynamic models in one of two ways, depending on the amount of data available per speaker.

For VOT and CSD, there is enough data to build models for each speaker which include both linguistic factors and time. This approach allows linguistic factors to have different effects for different housemates, as we saw was the case in the static models (most by-subject random slopes for linguistic factors were significant). The downside is that the estimated effects of linguistic factors may be less accurate than in the static model, since data from other housemates cannot be leveraged to estimate effects for which a particular housemate has little relevant data.

For vowels, the data is much sparser. There are fewer tokens per subject, and more degrees of freedom of linguistic factors to account for. Given the number of linguistic factors relative to the amount of data, it was not possible even in the static models to let the effects of linguistic factors vary by individual; the full dataset was needed just to effectively estimate group-level effects. Thus, it is not possible to model linguistic factors just using an *individual* housemate's data, nor to model time and linguistic factors simultaneously. the static models of the effects of linguistic factors on vowel formants. This choice means we are assuming all individuals have the same effects for linguistic factors—which is likely false—and we can only find time dependence in vowel formants that remains after these effects have been regressed out.

Model selection Model selection criteria are used for choosing among a set of candidate statistical models of the same data, particularly when they are non-nested (they cannot be ordered into a hierarchy, with the set of terms in each model a subset of the next-highest model). Different criteria measure the tradeoff between how well a given model fits the data and the model's size (i.e., number of parameters, data points) in different ways. Candidate models are then ranked by their values of the criterion, and either a single model is chosen (*model selection*) or the models are combined weighted by some function of their criterion values (*model averaging*) (Burnham and Anderson, 2002).

A variety of criteria exist, corresponding to different assumptions about the candidate models and what the purpose of model selection is. The most widely-used are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), or variants.[3] Which model selection criterion to use depends on the dataset being modeled. AIC is more appropriate in the 'tapering effects' scenario, where the response variable is a function of many different variables, with progressively smaller effect sizes, each of which will only be detectable once enough data is collected (Burnham and Anderson, 2004). BIC is more appropriate when there are relatively few large effects, and few or no smaller effects. Modeling phonetic variation falls squarely into the first category, so we use AIC.

However, it should be noted that which form of time dependence we choose below for different variables and housemates is generally not independent of the model selection criterion chosen. Using BIC gives much more conservative time dependence: almost no time trends (Types 3 or 4), more housemates with by-clip variation only, and many more housemates with no time dependence. This difference is not surprising, given that BIC tends to choose sparser models than AIC in many applications, in line with the different goals of AIC and BIC-based model selection (Burnham and Anderson, 2004). But it does

[3] Others are conditional AIC, Minimum Description Length, and the Deviance Information Criterion.

suggest that we are observing relatively small effects of time dependence, and must be cautious in generalizing from our results. The general issue of robustness of our results is discussed below (Sec. 7.6.1).

7.2 Variable 1: Voice onset time

7.2.1 Models

For each of the 12 speakers, we fit 12 linear mixed-effect models of the logarithm of automatically-measured VOT, as a function of both time and linguistic factors, and chose the best one according to AIC.

Each model has one of the 12 types of time dependence described above. By-clip variation corresponds to a by-clip random intercept, A linear time trend is simply a fixed-effect term for EPISODE. For nonlinear time trends (quadratic, cubic, and quartic), fixed-effect terms (2, 3, or 4) are included for restricted cubic splines with 3, 4, or 5 knots.[4]

The linguistic factors in each model are the same as in the static model for VOT (Sec. 6.1). Fixed-effect terms are included for speaking rate, height of the following vowel, consonant place of articulation, its interaction with vowel height, (log) word frequency, following segment type, and syllable stress; a by-word random intercept is also included, to account for variation among words not accounted for by the fixed effect, and the unbalanced distribution of tokens by word types. All fixed effects were centered (continuous variables) or Helmert-coded (factors) as in the static model.

For each speaker, we checked for signs of overfitting in the AIC-selected model, particularly near the endpoints, by comparing its predicted time trajectory with the empirical data. There was overfitting for one speaker (Stuart), whose 3 lowest-AIC models had cu-

[4] Restricted cubic splines with k knots are a basis for polynomial functions of order k-1 that are restricted to be linear near the endpoints (minimum and maximum EPISODE). The advantages of using restricted cubic splines rather than other bases for polynomial functions are discussed by Harrell (2001: 20–23); Stone and Koo (1985).

bic or quartic time trends and predicted VOT near 0 at the right endpoint. We used the lowest-AIC model which made sensible predictions (fourth-lowest AIC, quadratic time trend).

7.2.2 *Results*

Table 7.1 summarizes the form of time dependence selected for each speaker's dynamic VOT model: the standard deviation of the by-clip variation (if any) and the order of the time trend (if any) are shown. Speakers show diverse patterns of time dependence: two show none, five show by-clip variation, four show long-term trends, and one shows both by-clip variation and long-term trends. The table also shows the number of data points and the number of clips per housemate, to check if there is any relationship with which type of time dependence is selected. For example, it could be the case that we are only detecting time dependence for housemates with more data, making the fact that those housemates show time dependence an artifact of sample size rather than a meaningful individual difference.) No such relationship is evident.

7.2.2.1 Predicted time dependence

To visualize the dynamics of VOT in the house, we examine the time trajectory of VOT predicted by each housemate's model, with linguistic predictors held constant. Continuous predictors were set to their mean values, and each contrast for the discrete predictor (place of articulation) was set to zero. Thus, predicted values roughly describe the housemate's 'mean VOT'.

Individual speakers The solid lines in Fig. 7.1 show the predicted time trajectory for each housemate. For a housemate with by-clip variation, the line interpolates between the

Table 7.1: Summary of best model of time dependence for VOT for each speaker, chosen by AIC. σ_{clip} and Time Trend are the standard deviation of the model's by-clip random intercept and the degree of its long-term time trend, if any. N and N_{clip} are the number of data points and the number of clips per housemate.

Housemate	σ_{clip}	Time trend	N	N_{clip}
Dale	—	Linear	380	24
Darnell	0.1139	—	569	36
Kathreya	0.1044	—	286	24
Lisa	—	—	460	32
Luke	$3{\cdot}10^{-7}$	—	652	28
Michael	0.0706	—	834	44
Mohamed	—	—	409	28
Rachel	—	Linear	392	35
Rebecca	—	Linear	181	17
Rex	0.0421	Linear	615	38
Sara	0.2312	—	221	16
Stuart	—	Quadratic	202	17

model-predicted value for each clip.[5] For housemates without by-clip variation (Types 1, 3), the line simply shows the predicted time trend; if there is no time dependence, the line is flat.

We can evaluate the predicted time trends by comparing to the empirical trends in each housemate's data. To get a sense both of each housemate's empirical trend for VOT, as well as how much VOT variation he shows on a given day, we compute the mean VOT for each word produced at least once on that day. Each dashed line in Fig. 7.1 shows the LOESS-smoothed time trajectory of these mean VOTs for one housemate, with shading indicating 95% confidence intervals. The notches on the x-axis show when clips occur. fdsf In comparing a speaker's predicted and empirical time trajectories, it should be kept in mind that in general the height (=position on the VOT axis) of the two will be different. The predicted trajectory's height depends on the particular values we are holding linguistic factors constant at, which will not usually correspond to the mean over all words pro-

[5] The BLUP for the clip's random intercept, plus the predicted VOT on the day when the clip occurs (global intercept, plus time trend if present).

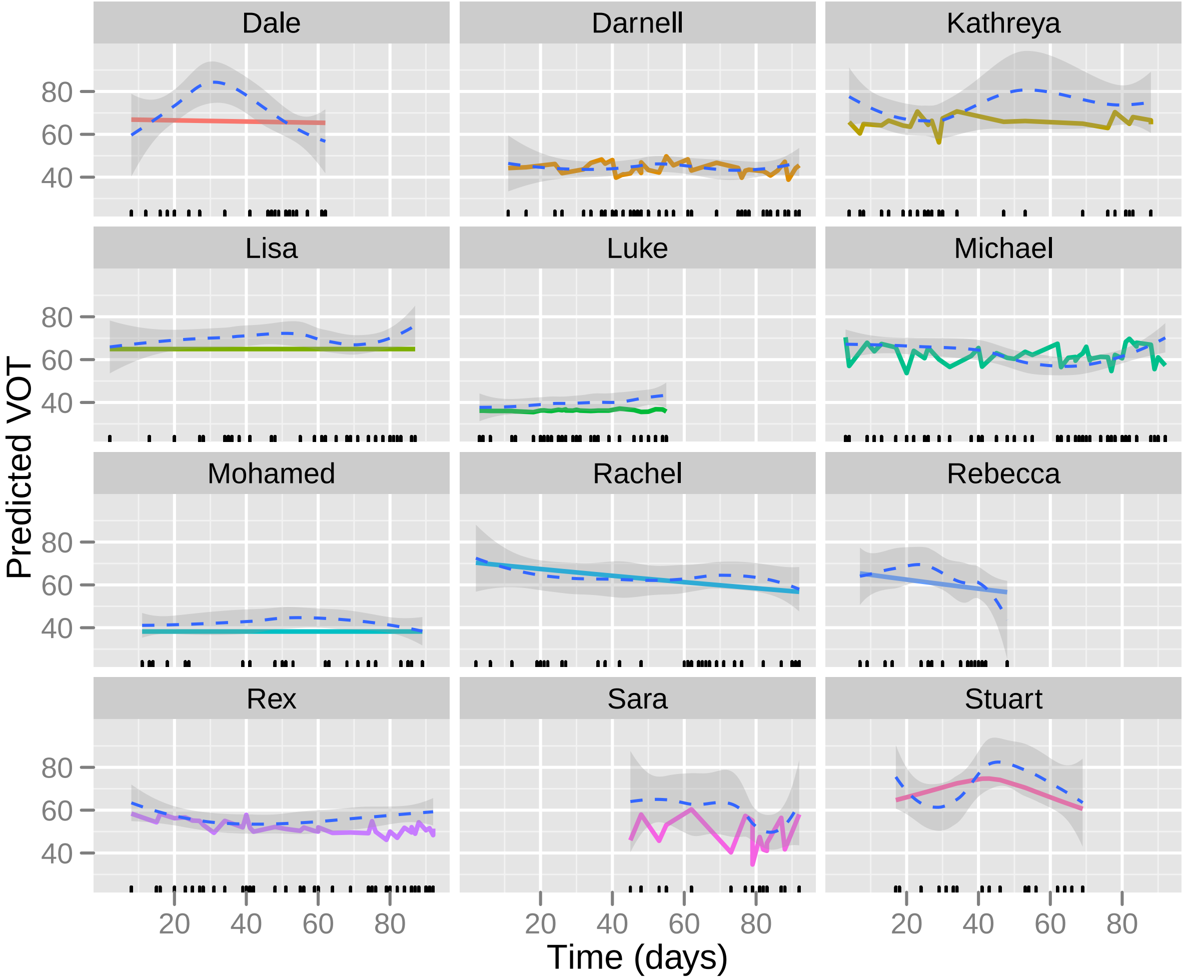

Figure 7.1: Solid lines: Best model of VOT time dependence for each speaker (see Table 7.1), with all linguistic predictors held constant. Non-smooth lines indicate by-clip variation, with linear interpolation between each clip's fitted value. Dashed lines and shading: Empirical time trajectory of speaker's VOT, LOESS-smoothed over mean values for each word on each day, with 95% confidence intervals.

duced on a given day. However, the predicted and empirical patterns of time dependence (independent of height) will generally be comparable.[6] In general, for all time trajectories presented in this chapter, the dynamic models predict *relative* time dependence (the difference between the variable's value at two time points), and the height of the trajectory depends on how linguistic factors are controlled for.

With this caveat, the models generally fit the observed dynamics well. They are also conservative, in the sense of predicting less change than is empirically observed. For speakers where time trends are predicted (Rachel, Rebecca, Rex, Stuart), the predicted trend is smoother than the empirical trend. For most speakers where no time dependence is predicted, there is little empirical time trend. The exception is Dale, for whom no time trend is predicted despite an apparent empirical trend. For most speakers where by-clip variation is predicted, the estimated by-clip variance (height of squiggles) is similar to the observed by-clip variation among words (height of shading). The one exception is Kathreya, where the mismatch can be attributed to very sparse data (few clips) in the region where the fit is poor.

All speakers Fig. 7.2 shows predicted time trends for all speakers together, with shading now indicating by-clip variation. There no clear pattern of overall convergence. However, most speakers who show time trends start with high VOT (Dale, Rachel, Rebecca, Rex), and decrease over the season; Stuart also begins with high VOT, increases, and then decreases. Interestingly, the mean VOT in the house (across speakers, where each speaker's value on a given day is calculated as the mean of word means) is between 55 and 60 ms over the course of the season. Rex could be seen as shifting away from this value, and the other four speakers towards it.

[6] Assuming the speaker produces a representative sample of words on each day; otherwise the mean VOT will be skewed by properties of the words.

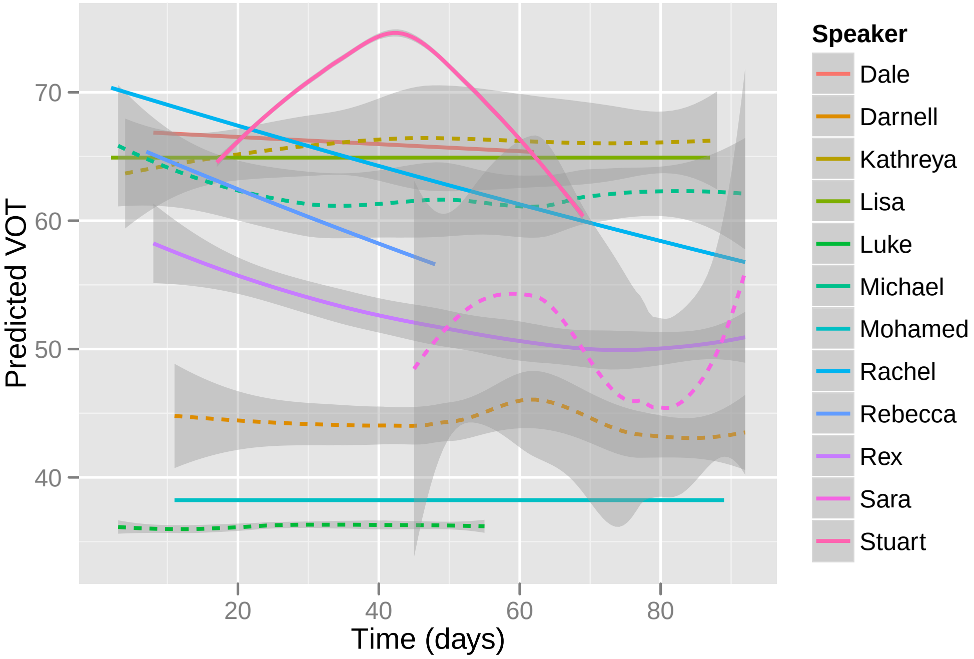

Figure 7.2: Best models of VOT time dependence for each speaker, with linguistic predictors held constant. Lines with shading are LOESS-smoothed trajectories for speakers with by-clip variation, computed over predicted values for each clip, with 95% confidence intervals. Dotted lines are trajectories for housemates with only by-clip variation.

7.3 Variable 2: Coronal stop deletion

7.3.1 Models

As in the static model of CSD, we exclude one speaker (Katherya) who deletes near-categorically. For the other 11 speakers, we fit 12 mixed-effects logistic regression models of final coronal realization as a function of time and linguistic factors, and choose the best one using AIC.

The 12 types of time dependence are the same as for VOT. Most linguistic factors are the same as in the static model of CSD. Fixed-effect terms are included for preceding context, following context, (log) word frequency, and following neutralizing context; a by-word random intercept is also included, to account for variation among words not accounted for by the fixed effect, and the unbalanced distribution of tokens by word types.

In the static model, we included morphological class as a factor with three levels (monomorphemes, semi-weak pasts, weak pasts). For the dynamic models, we only include a two-way distinction (monomorphemes vs. pasts), because some of the 12 speakers have very few semi-weak past tokens, which causes problems with model identifiability.[7] Coding morphological class as monomorphemes vs. bimorphemes is in fact relatively common in studies of CSD, since semi-weak pasts are so infrequent. All fixed effects were centered (continuous predictors) or Helmert-coded (factors), as in the static model.

For each speaker, we checked for signs of overfitting in the AIC-selected model, particularly near the endpoints, by comparing its predicted time trajectory with the empirical data. For one speaker (Lisa), the best three models had cubic or quartic time trends where the predicted CSD rate blows up at the endpoints. We used the next-best model for Lisa which made sensible predictions (fourth-lowest AIC, quadratic time trend).

[7] For these speakers, morphological class is highly collinear with the intercept, since the value of the intercept is always 1, and the value of the "is this a semi-weak past" predictor is nearly always 0.

Table 7.2: Summary of best model of time dependence of CSD rate for each speaker, chosen by AIC. σ_{clip} and Time Trend are the standard deviation of the model's by-clip random intercept and the degree of its long-term time trend, if any. N and N_{clip} are the number of data points and the number of clips per housemate.

Speaker	σ_{clip}	Time trend	N	N_{clip}
Dale	—	—	293	24
Darnell	—	—	737	39
Lisa	—	Quadratic	492	30
Luke	—	—	568	28
Michael	0.596	—	777	41
Mohamed	—	—	288	24
Rachel	—	—	486	38
Rebecca	—	Quadratic	231	16
Rex	—	—	586	28
Sara	—	—	262	16
Stuart	—	—	262	17

7.3.2 *Results*

Table 7.2 summarizes the type of time dependence in each speaker's dynamic model of CSD rate, in the same format as Table 7.1. Eight speakers show no time dependence, one shows by-clip variation, and two show time trends. There is no systematic relationship between the amount of data (N or N_{clip}) and the type of time dependence selected, across speakers.

7.3.2.1 Predicted time dependence

As for VOT, we examine the time trajectory predicted for each speaker's CSD rate when all linguistic factors are held constant: continuous predictors at their mean value, and each contrast for the discrete predictors at 0. The predicted values roughly give a housemate's 'mean CSD rate', in logit space.

Individual speakers The solid lines in Fig. 7.3 show the predicted time trajectories for each housemate, in the same format as used as for VOT (Fig. 7.1): jagged/smooth lines indicate the presence/absence of by-clip variation.

To visualize empirical trends in each housemate's data, we applied a generalized additive model smoother with a logistic link to the empirical CSD rates for each word produced at least once by a speaker on a given day.[8] Each dashed line in Fig. 7.3 shows a housemate's estimated time trajectory, with shading indicating 95% confidence intervals.

Again, the height of each speaker's empirical and predicted trajectories are not comparable, but their shapes are. The empirical trajectories are higher than the predicted trajectories because they do not incorporate by-word offsets in deletion rate (as the by-word random intercept does, in the models underlying the predicted trajectories). Thus, very frequent words such as *and* and *just*, which have high deletion rates, are over-weighted.

With this caveat, the models fit the observed trajectories reasonably well. Most speakers have nearly-flat empirical trajectories, and no predicted time trend. For one speaker (Rebecca) with a predicted time trend, there is a good fit with the empirical trend. For the other speaker (Lisa) with a predicted time trend, the fit with the empirical trend is poor—the model predicts too much change. Closer inspection of Lisa's data shows that her model is overfitting to data from her first clip, which happens to have a high deletion rate. Because this clip is so much earlier than her others, it has disproportionate influence when fitting a polynomial time trend, the only kind we are allowing. Lisa's CSD model is the one case where our method predicts markedly more change than is observed in the empirical data.

All speakers Fig. 7.4 shows the predicted time dependence of CSD rate for all speakers, in probability space (=the inverse logit of the model's prediction). Only three speakers show any time dependence, and there is no obvious pattern of convergence. Interestingly, the only speaker who shows by-clip variation is Michael, the one Scottish speaker. Sociolinguistic studies of CSD have generally found very little style shifting; an important

[8] Using the defaults for `gam` in the `mgcv` package in `R`, where the smoothing parameter is estimated using generalized cross validation (Wood, 2012).

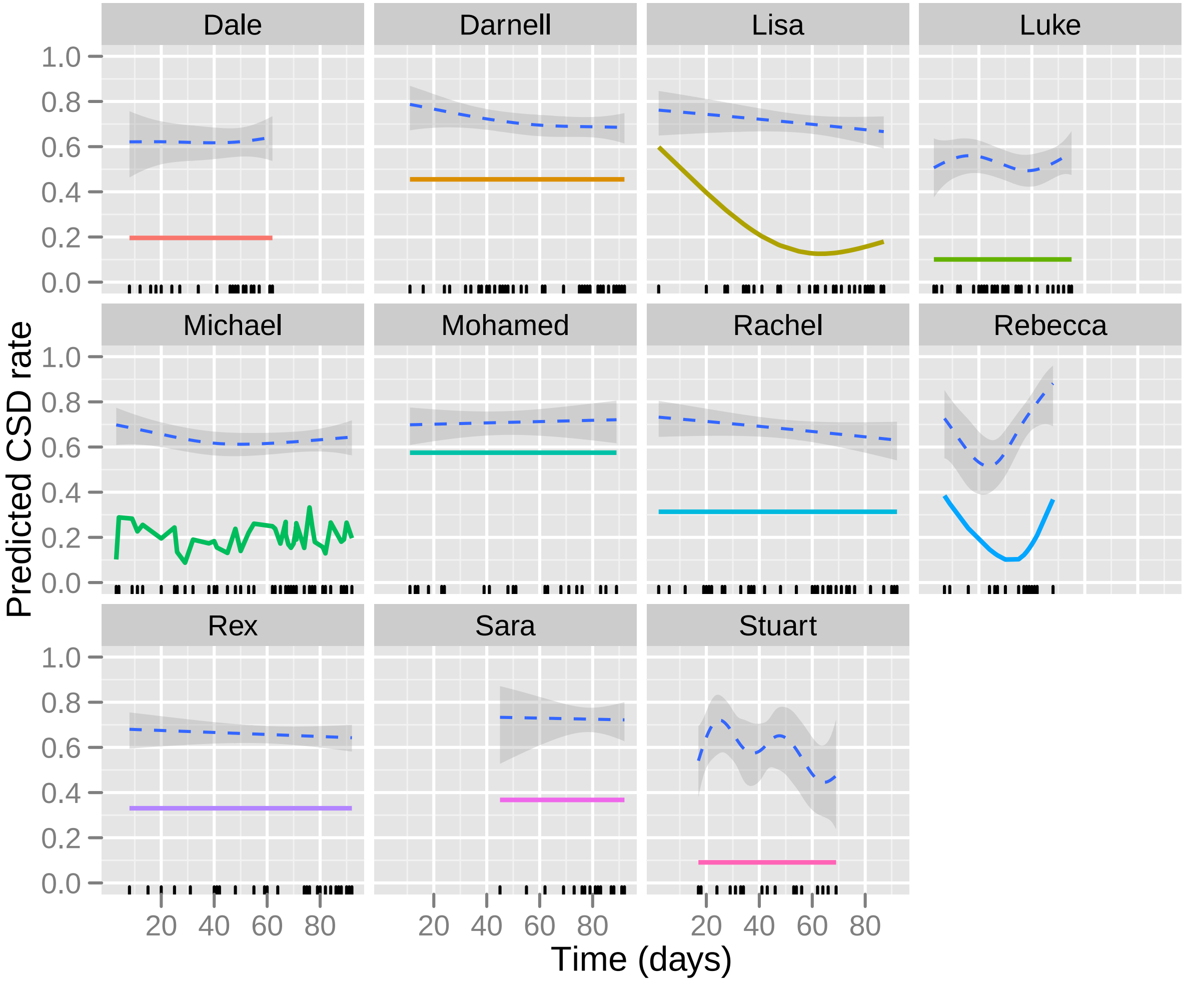

Figure 7.3: Solid lines: Best model of CSD rate time dependence for each speaker, with linguistic predictors held constant. Non-smooth lines indicate by-clip variation, with linear interpolation between each clip's fitted value. Dashed lines and shading: Empirical time trajectory of speaker's CSD rate, smoothed using a generalized additive model (logistic link) on proportions of CSD realization for each word type in each clip, with 95% confidence intervals.

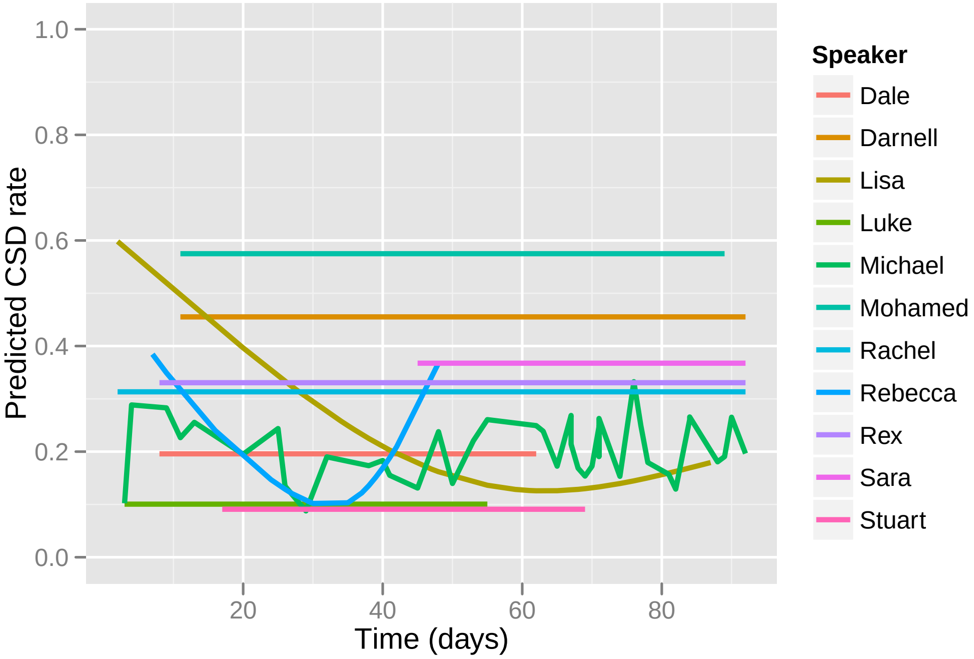

Figure 7.4: Predicted time dependence of CSD rate under the best model for each speaker, with linguistic predictors held constant. Smooth trajectories are for speakers with no by-clip variation. The non-smooth trajectory indicates by-clip variation, with linear interpolation between each clip's fitted value.

exception is Smith et al.'s study of Scottish adults and children (Smith et al., 2009). It is tempting to speculate that Michael's by-clip variation is due to style shifting.

Compared to other variables, the number of speakers who show no time dependence is striking. A possibility worth considering is that more speakers in fact show by-clip variation which is not detected, because it is impossible to determine how much the deletion rate for some clips differs from the norm: clips containing only a few tokens, or only extremely frequent word types, ('and', 'just'), or both. Mixed models are supposed to be robust to such unbalanced data (i.e., clips containing extremely different numbers of tokens or word types). But of course, for sufficiently unbalanced data no by-clip variation will be found even if it is underlyingly present.[9] As a quick test of whether this might be responsible for the lack of predicted time dependence, we refit all models using data

[9] We thank Gregory Guy for pointing this out.

only from clips with at least 10 tokens and five types. The best model for each speaker using this restricted dataset looks similar to the model resulting from using the full dataset, except for Michael, who now shows *no* time dependence. Thus, slightly *less* time dependence is predicted when uninformative clips are excluded, suggesting that they are not responsible for the lack of predicted time dependence.

7.4 Variables 3–5: Vowel formants

7.4.1 Models

As discussed above, a different procedure from VOT and CSD is used to build dynamic models for vowel formants, due to data sparsity. For each speaker, we model formant residuals from the six static models: F1 and F2, for GOOSE, STRUT, and TRAP′. For each of the six cases, we try 12 models corresponding to the different types of time dependence. In addition, each model includes a by-word random intercept, to control for by-word pronunciation variation within a housemate's dialect (beyond the effects of consonantal context) and for the unbalanced distribution of tokens by word types.

Some of the 12 speakers (who are in the house for at least 50 episodes) have very few tokens for one or more of the vowels, making it impossible to distinguish between by-clip variation and long-term trends. We therefore choose an arbitrary cutoff, and only model speakers who have at least 90 tokens for a given vowel (see Table 7.3).

7.4.2 Results

Table 7.3 summarizes the form of time dependence selected for each speaker, for F1 and F2 residuals of each vowel, in the same format as Tables 7.1 and 7.2. Five of eight speakers show some time dependence (in F1 or F2) for GOOSE, eight of ten for STRUT, and all eight for TRAP′. Thus, the realization of TRAP′ is the most consistently plastic variable within speakers among the five variables we have examined. There is much more by-clip variation in F1 than in F2, across vowels. For both formants for each vowel, there does not appear to be any systematic relationship between the amount of data for a speaker (N or N_{clip}) and the type of time dependence in their dynamic model.

7.4.2.1 Predicted time dependence

Because the static models for each vowel's formants included a by-speaker random intercept, the dynamic models all make predictions centered at 0 (because they model residuals from the static models). To compare speakers' predicted formant trajectories in vowel space, we need to add constants to each speaker's formant residual trajectories so they are centered at her mean formant values for each vowel. We do so by adding each speaker's random intercepts from the six static models to the predictions of her six dynamic models. The resulting predictions can be interpreted roughly as 'mean normalized formant values'.[10]

Individual speakers Figures 7.5–7.7 show predicted time trajectories for each formant (solid lines). Also shown are empirical time trajectories, smoothed over all tokens (dashed lines), and 95% confidence intervals (shading). Overall, the predicted trajectories match the empirical trends quite well. The dynamic models are generally conservative, in the sense of predicting less change than is empirically observed, as was also the case for VOT. When a time trend is predicted, it is nearly always shallower than the corresponding empirical trend. In most cases where by-clip variation is predicted, the estimated by-clip variability (height of squiggles) is similar to the observed by-token variability (height of shading). In some cases no time dependence is predicted (e.g., GOOSE F2 for Darnell), even though a trend seems possible from the empirical data.

[10] For VOT and CSD, we held linguistic factors constant to obtain time trajectories. This step is not necessary for the predictions of the dynamic models for formant residuals, since the effects of linguistic factors have already been regressed out.

Table 7.3: Summary of best models of time dependence for F1 and F2 residuals for each vowel, chosen by AIC. Only models for speakers with at least 90 tokens of a vowel are shown. Residuals are from the static models in Sec. 6.3 (see text). σ_{clip} is the standard deviation of the model's by-clip random intercept; the value in Hz is given in parentheses. Time Trend is the degree of the model's long-term time trend, if any. N and N_{clip} are the number of data points and the number of clips per housemate.

Vowel	Speaker	F1		F2		N	N_{clip}
		σ_{clip}	Time trend	σ_{clip}	Time trend		
GOOSE	Darnell	—	Linear	—	—	205	37
	Kathreya	—	—	—	—	111	23
	Lisa	—	—	—	—	126	31
	Luke	0.113 (20)	—	0.149 (68)	—	185	28
	Michael	0.129 (19)	Linear	—	—	192	41
	Mohamed	0.158 (21)	—	—	—	103	24
	Rachel	0.139 (27)	—	—	—	133	35
	Rex	—	—	—	—	242	37
STRUT	Dale	—	—	—	—	110	23
	Darnell	0.077 (10)	—	—	Linear	284	36
	Kathreya	0.242 (54)	—	—	—	189	25
	Lisa	0.130 (25)	—	—	Linear	234	30
	Luke	—	—	—	—	286	28
	Michael	0.114 (17)	—	—	—	451	45
	Mohamed	0.316 (43)	—	0.131 (43)	—	161	27
	Rachel	0.173 (33)	—	0.108 (48)	—	189	33
	Rex	0.137 (15)	—	—	—	311	38
	Sara	—	Quadratic	—	Quadratic	144	15
TRAP′	Darnell	0.153 (20)	—	—	Quadratic	271	37
	Kathreya	—	—	—	Linear	92	23
	Lisa	0.146 (52)	Linear	—	—	149	31
	Luke	0.246 (44)	—	—	—	202	28
	Michael	0.102 (15)	Linear	0.075 (31)	—	337	41
	Rachel	—	Linear	0.260 (115)	—	94	32
	Rex	0.190 (20)	—	—	—	123	34
	Sara	0.335 (62)	—	—	Linear	107	18

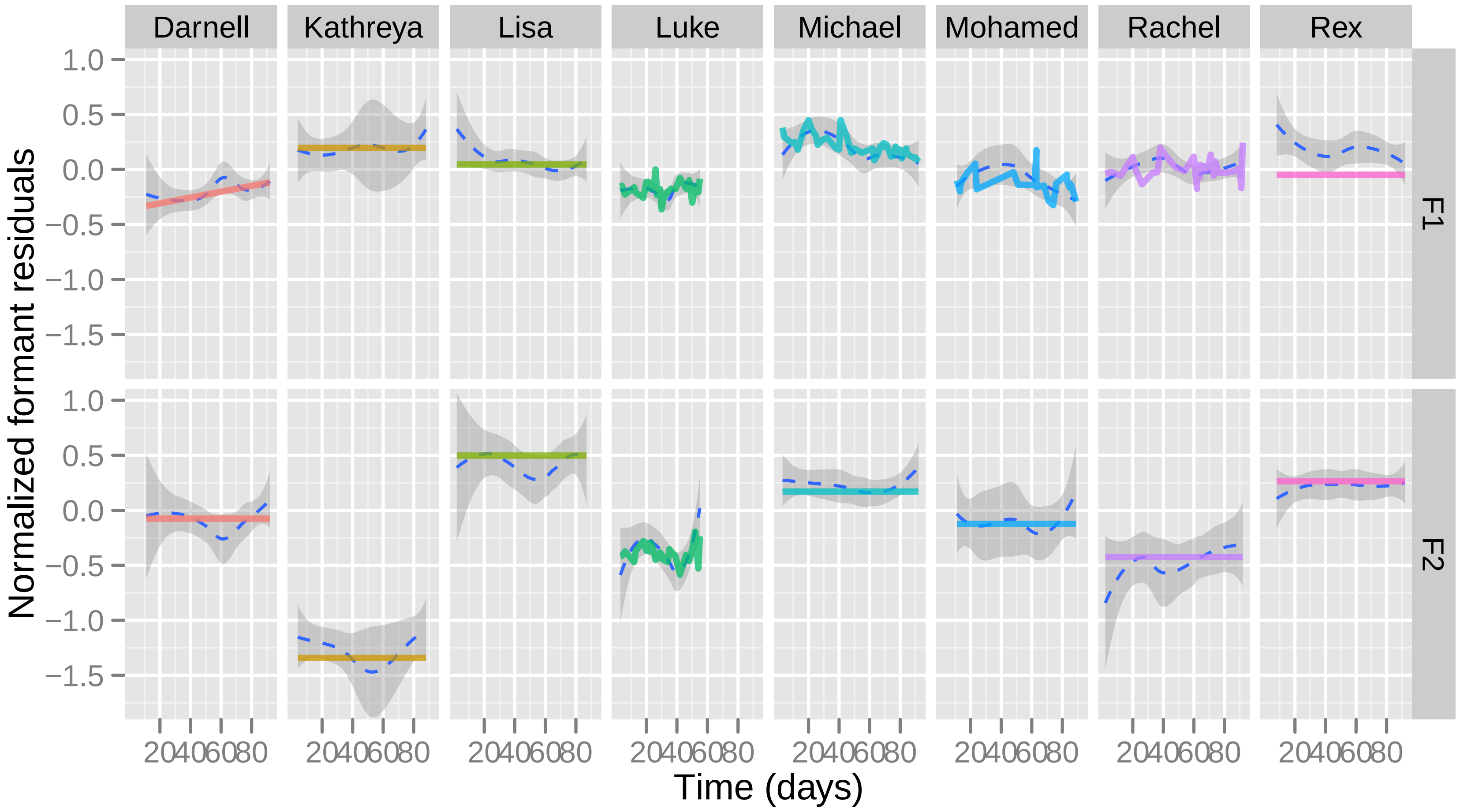

Figure 7.5: Solid lines: Predicted time trajectories for GOOSE F1 and F2 residuals for speakers with at least 90 tokens. Non-smooth lines indicate by-clip variation, with linear interpolation between each clip's fitted value. Dashed lines and shading: LOESS-smoothed empirical time trajectory of speaker's F1 and F2 residuals, with 95% confidence intervals.

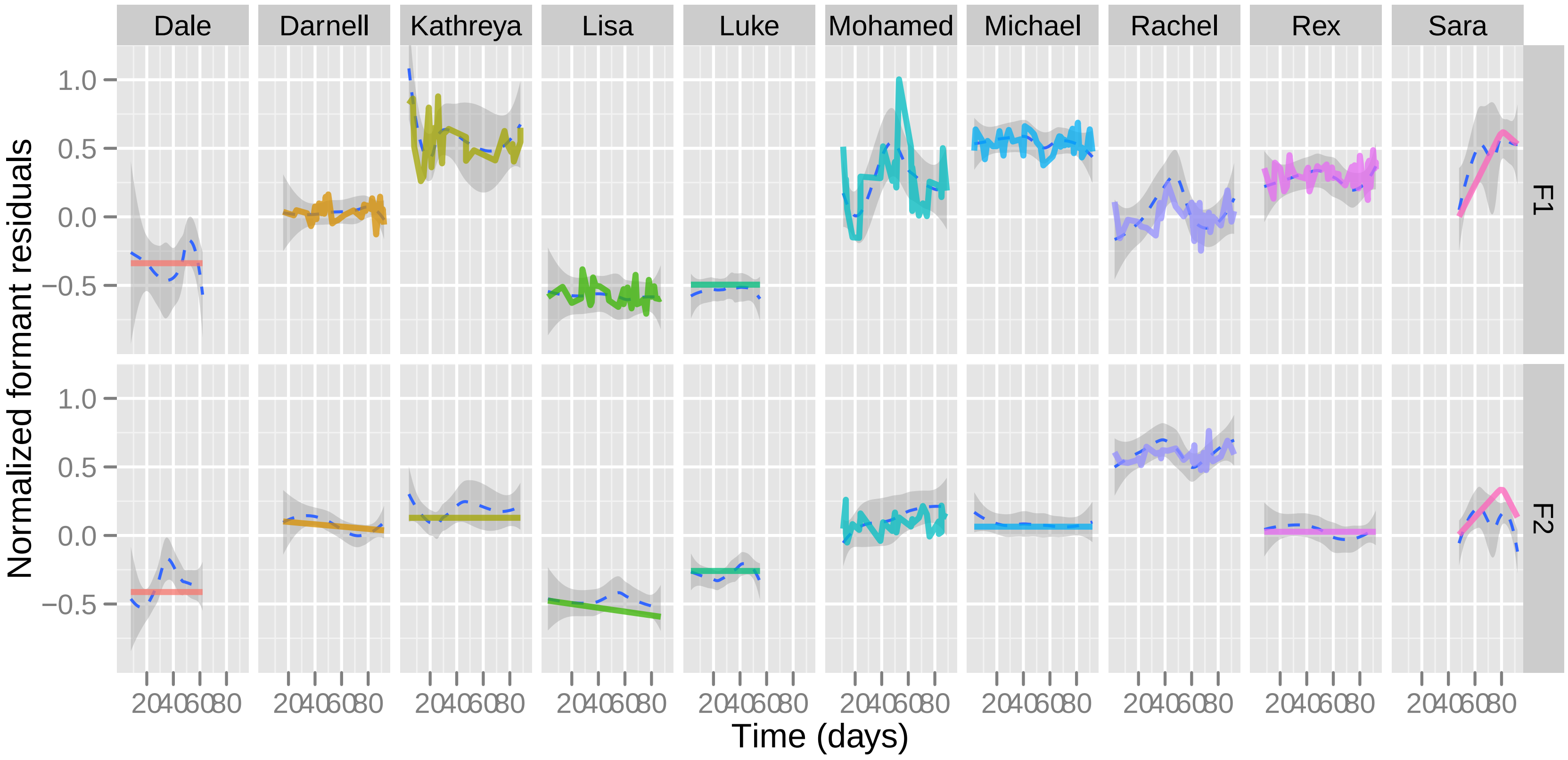

Figure 7.6: Solid lines: Predicted time trajectories for STRUT F1 and F2 residuals for speakers with at least 90 tokens. Non-smooth lines indicate by-clip variation, with linear interpolation between each clip's fitted value. Dashed lines and shading: LOESS-smoothed empirical time trajectory of speaker's F1 and F2 residuals, with 95% confidence intervals.

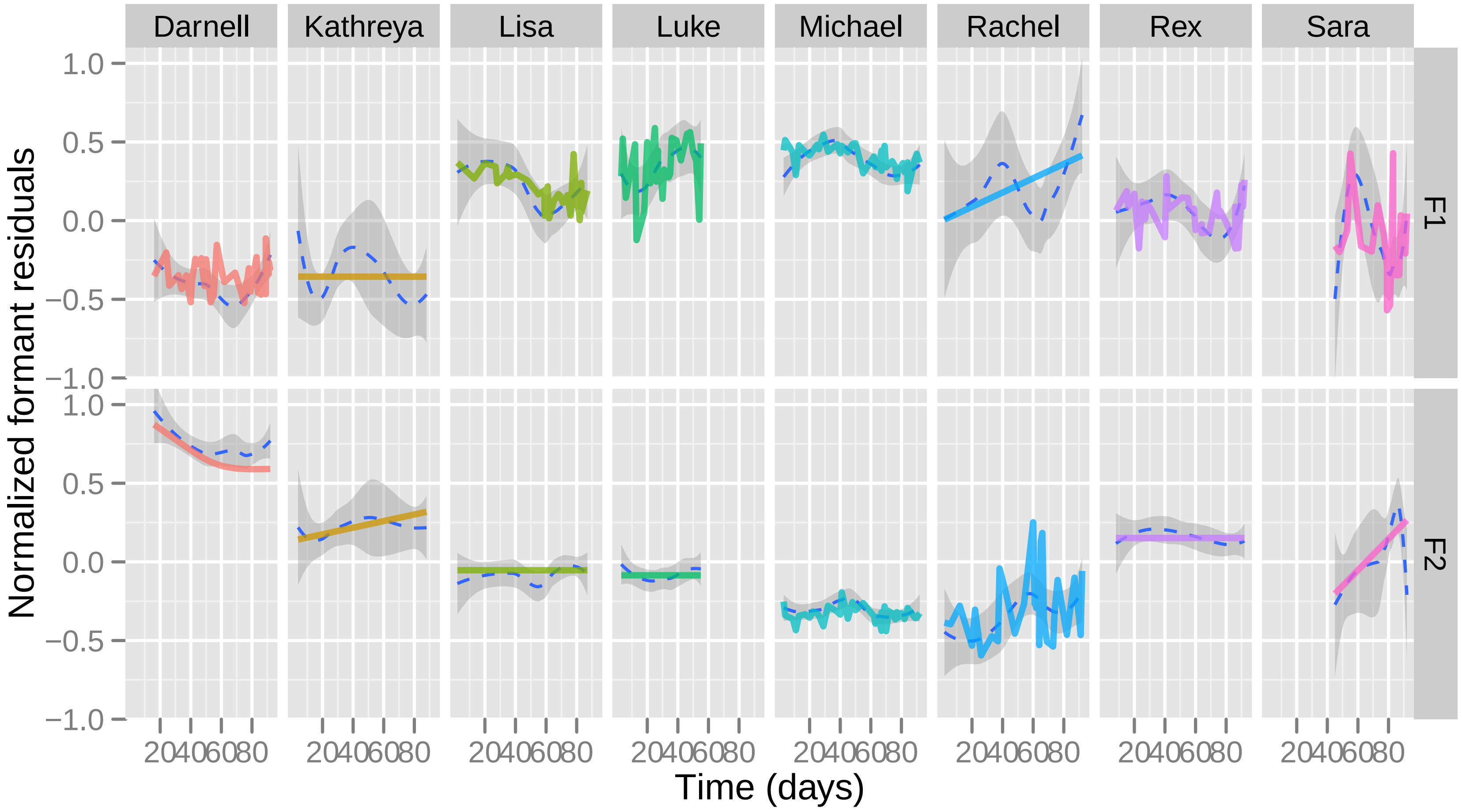

Figure 7.7: Solid lines: Predicted time trajectories for TRAP′ F1 and F2 residuals for speakers with at least 90 tokens. Non-smooth lines indicate by-clip variation, with linear interpolation between each clip's fitted value. Dashed lines and shading: LOESS-smoothed empirical time trajectory of speaker's F1 and F2 residuals, with 95% confidence intervals.

Table 7.4: Number of speakers who show each type of time dependence for each variable.

Time Dependence	VOT	CSD	GOOSE F1	GOOSE F2	STRUT F1	STRUT F2	TRAP′ F1	TRAP′ F2
None	2	8	3	7	2	5	1	3
By-clip	5	1	3	1	7	2	4	2
Time trend	4	2	1	0	1	3	1	3
Both	1	0	1	0	0	0	2	0

All speakers Fig. 7.8 shows the predicted time trends in F1 and F2 for each vowel for all speakers together, with shading indicating by-clip variation.

GOOSE shows the clearest case of possible convergence, in F1. The two speakers with time trends, Michael and Darnell, start out with the highest and lowest F1, and over the course of the season move towards the mean F1 value in the house. F2 is remarkably flat, with only one speaker showing by-clip variation. For STRUT, three speakers show time dependence in F1 or F2, but there is no pattern that can be interpreted as convergence. For TRAP′, three speakers show time dependence in each formant, and some could be argued to be converging towards the mean value in the house. For F1, the two speakers who start highest (Michael, Lisa), decrease over the season; for F2, the speaker who starts highest (Darnell) also decreases. But the other three speakers cannot be interpreted in this way: Rachel moves away from the mean for F1, as do Katherya and Sara for F2.

7.5 Variability

Fig. 7.9 shows the predicted patterns of time dependence for all variables, for all speakers. The distribution of types of time dependence among different variables is summarized in Table 7.4.

There is huge variability in the type of time dependence across speakers and variables. Usually, different speakers show qualitatively very different time dependence for a given variable; within the same speaker, different variables usually have qualitatively

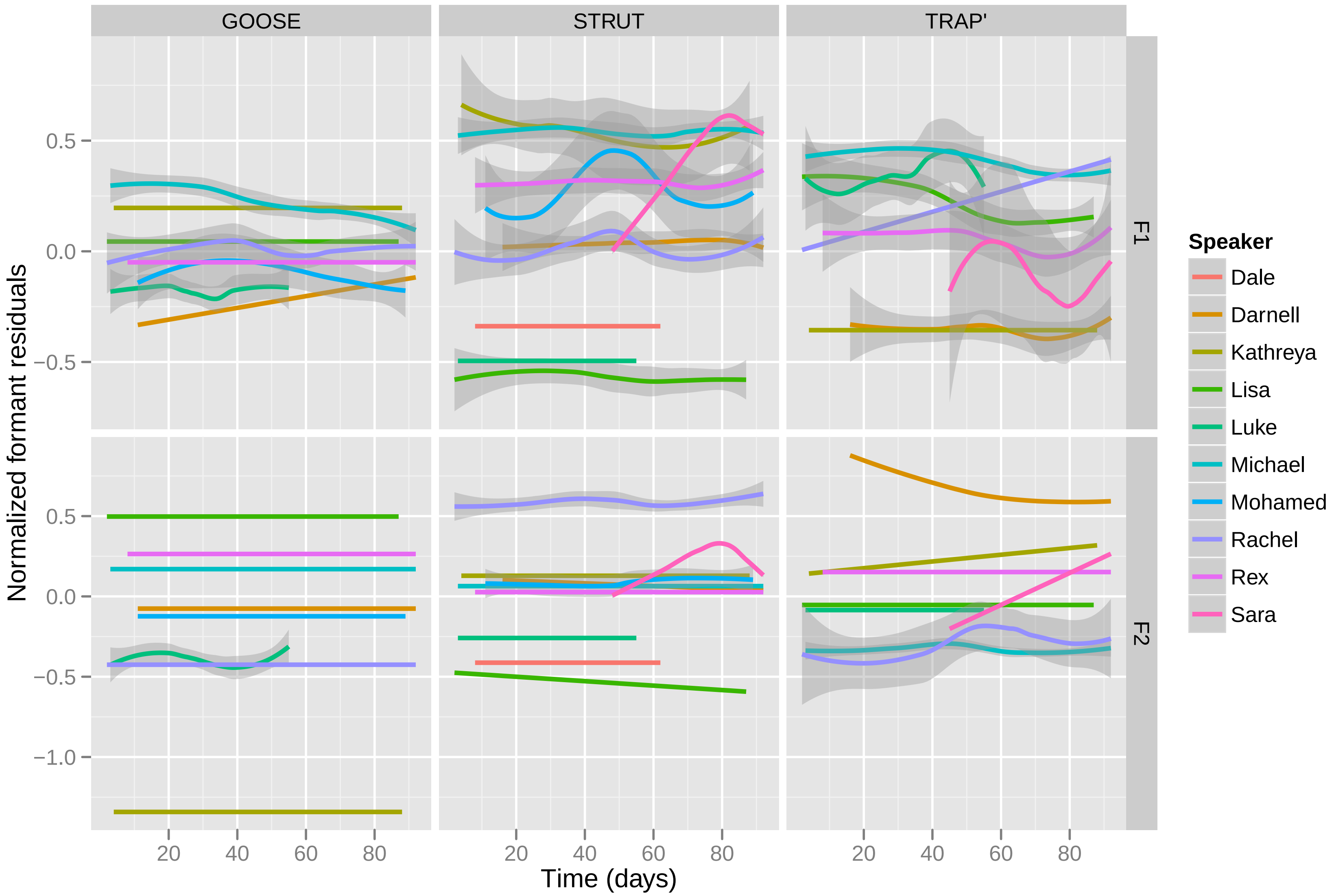

Figure 7.8: Predicted time dependence for F1 and F2 residuals for GOOSE, STRUT, and TRAP′, for speakers with at least 90 tokens. Lines with shading are LOESS-smoothed trajectories for speakers with by-clip variation, computed over predicted values for each clip, with 95% confidence intervals.

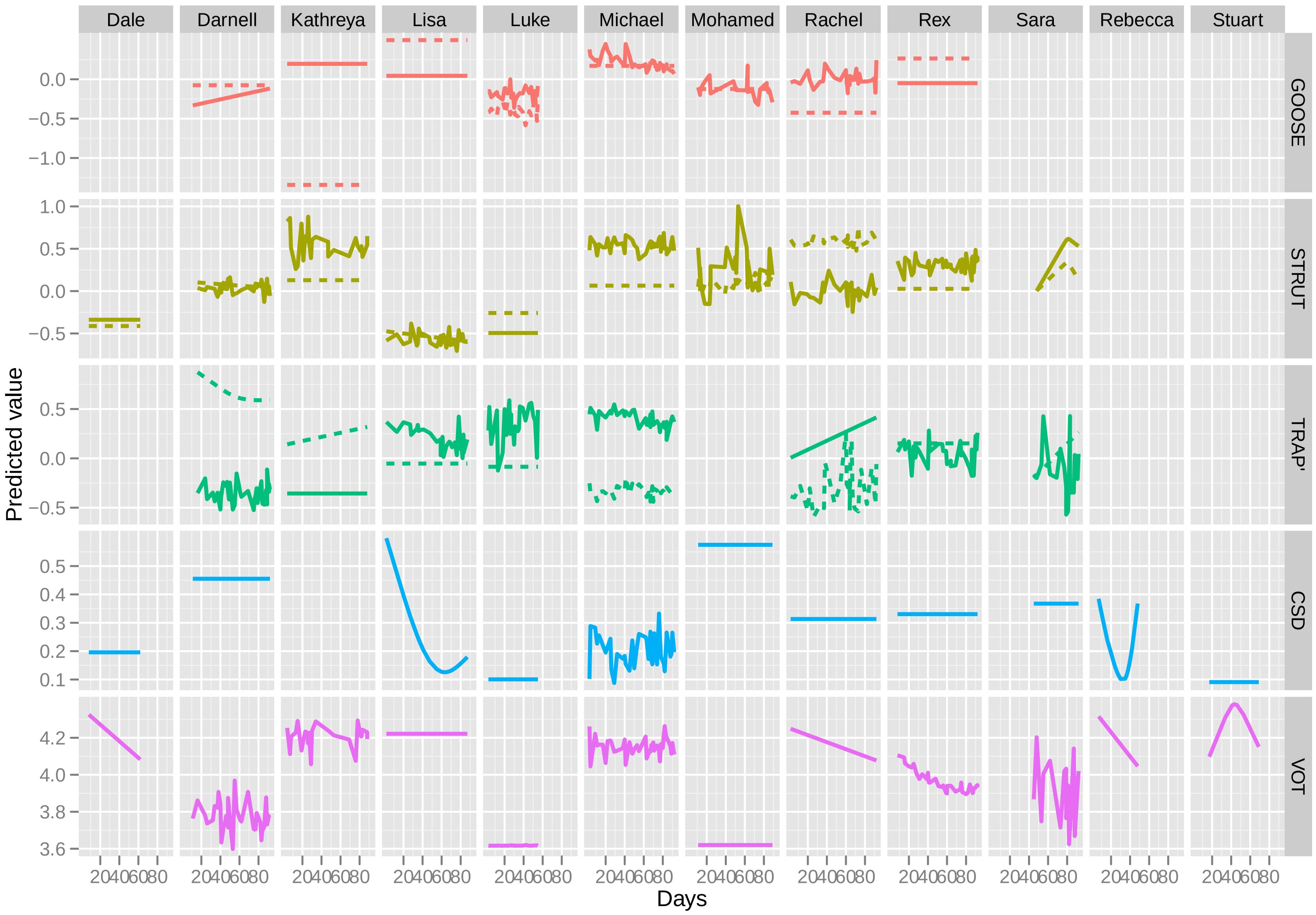

Figure 7.9: Predicted time dependence for all variables, for all speakers. Non-smooth lines indicate by-clip variation, with linear interpolation between each clip's fitted value.

different dynamics. This is one of our main findings: *variability is the norm* for time trajectories of phonetic and phonological variables in this medium-term data. In this sense, the dynamics of variables in the house look more like long-term change (highly variable outcomes for different speakers and variables) than short-term change (relatively robust shifts across speakers and variables).

However, there is also some evidence for structure in the variability, at the level of clips, speakers, and variables.

7.5.1 By-clip variability

Many of the dynamic models show significant by-clip variability in a speaker's use of a variable. By-clip variability could be due to random day-to-day fluctuations, or to style-shifting (any variation due to the social or situational context). We consider two possible sources of style shifting.

Conversational topic and context Housemates come to the diary room for a variety of reasons. How they use a variable within a clip could depend on why they are in the diary room, or the topic of conversation. As a rough test for such dependence, we coded clips as one of six types, depending on why the housemate was in the diary room.[11] Only one type had sufficient sample size (number of clips) to check for an effect: whether a housemate was called to the diary room to nominate or not. Recall that once per week, housemates are called to the diary room to nominate two other housemates for eviction, and to explain their decision. Nominations are a very socially-salient event in the house, and housemates presumably are aware that footage of them nominating is very likely to be broadcast. At least impressionistically, some housemates seem to talk in a different register when nominating.

[11] Housemates could be called to the diary room to perform a task, to nominate, to be disciplined, or for another reason (four types); they could also come on their own initiative (one type), or it could be unclear why they are in the diary room (one type).

To test for style shifting when talking about nominations, we use a new by-clip variable, NOMINATION, which is 1 for tokens in a nomination clip and 0 otherwise. As a clip level variable, NOMINATION can be thought of as a third type of time dependence (along with random by-clip variation and time trends). We add two new models of time dependence to the 12 already tried: one including just a fixed effect of NOMINATION (Model 13), and one including a fixed effect of NOMINATION and a by-clip random intercept (Model 14). In the first model, a housemate's use of a variable only differs as a function of whether she is nominating; in the second model, her use of a variable varies both randomly among clips, as well as by whether she is nominating. For each variable, we now choose a best model using AIC from among 14 models.

Table 7.5 summarizes the effect of adding these two models on which types of time dependence are selected for speakers for each variable. In a large majority of cases, there is no change: the type of time dependence chosen is one of the original 12. But there are three types of cases where the best model now involves NOMINATION. In six cases, Model 14 is chosen instead of by-clip variation, meaning that some random by-clip variation can instead be attributed to whether the speaker is nominating or not. In two cases, Model 13 is chosen instead of by-clip variation, meaning that *all* by-clip variation can be attributed to whether the speaker is nominating. In five cases, Model 13 is chosen instead of no time dependence, meaning the speaker's usage does differ among clips, just not randomly. Importantly, we do not see a fourth possible type of change, where what we thought was a time trend turns out to be attributable to whether the speaker is nominating.

The 12 cases where the best model now involves NOMINATION provide good evidence for at least some style shifting. However, most by-clip variation remains unaccounted for.

Big Brother identity Recall that in each clip housemates are speaking to one of many speakers with very different accents, all called simply Big Brother. Although Big Brother usually says little, it is possible that housemates are influenced by his or her speech. Ide-

Table 7.5: Summary of the changes in types of time dependence chosen for each speaker for each variable, when two types of time dependence including NOMINATION are added to the set of candidates: "nomination only" and "by-clip plus nomination." See text.

	VOT	CSD	GOOSE		STRUT		TRAP′	
			F1	F2	F1	F2	F1	F2
No change	10	12	8	5	7	8	6	6
By-clip only→by-clip + nomination	1	0	0	0	2	1	2	1
None→nomination only	0	0	0	3	0	1	0	1
By-clip only→nomination only	1	0	0	0	1	0	0	0

ally we would test for this possibility by determining which clips correspond to which Big Brother, and check for any systematic shifts as a function of Big Brother's identity. We tried to perform the first step of this process by simply listening to Big Brother's speech in different clips, and found it very difficult to determine which clips have the same Big Brother, given how little he or she says per clip. In future work we will try other methods, such as having British speakers judge similarity among the (British) Big Brothers in different clips, or quantifying similarity using speaker-identification methods from the automatic speech recognition literature.

Importantly, though, whatever accommodation to Big Brother occurs cannot be too widespread, or we would expect to see by-clip variation in *every* case. In the many cases of time trends without by-clip variation, or no time dependence, any accommodation that occurred was not extensive enough to be included in the best model. Under the (questionable) assumption that roughly similar levels of accommodation will occur for all speakers and for all variables, the lack of by-clip variation in many cases suggests that accommodation is not a significant contributor to cases where by-clip variation is observed. Nonetheless, checking for accommodation to Big Brother is a crucial direction for future work.

7.5.2 Dynamics of different speakers

Given time trajectories for multiple variables for each speaker, a natural question is whether different speakers show characteristic patterns of time dependence, across variables. Qualitative examination of Fig. 7.9 suggests that they may. Some speakers seem less susceptible to change: Luke and Mohamed show no time trends for any variable, and by-clip variation for only some variables, while Michael shows by-clip variation for all variables, and only very slight time trends.[12] Other speakers seem more susceptible to change: both Rachel and Sara show large time trends or large by-clip variation for all but one variable.

These differences between speakers suggest an interesting hypothesis, that adults differ in the plasticity of their sound system. Of course, our classification of speakers as more or less plastic is post-hoc. The plasticity hypothesis must be tested in future work, either by investigating either more variables for these 12 speakers, or using a different dataset containing more speakers.

Why might different speakers show different patterns? The same trouble we had in explaining by-speaker differences in the static models is again relevant: in the absence of a large difference between speakers with an obvious interpretation, it is hard to say much with 12 speakers. No patterns jump out as relatable to factors from the short-term or long-term literature which impact how much speakers change. (For example, age of arrival, length of stay, and identity factors in dialect change studies; social or attitudinal factors in short-term studies.) Cases of possible convergence are discussed below, under by-variable differences.

Speakers are constantly interacting in the house, and the type and frequency of interactions between particular speakers is strongly impacted by socially-salient events, like the division of housemates into two groups (“heaven” and “hell”) for one month.

Given the extreme amount of social interaction in the house, we might expect speak-

[12] We also have the most data for Michael, so it is possible we are able to resolve small time trends for him, but not speakers with less data.

Table 7.6: Percentage of speakers for each variable who show by-clip variation, a time trend, or either pattern, and whether possible convergence is observed.

	VOT	CSD	GOOSE	STRUT	TRAP′
By-clip variation (with or without time trend)	41%	9%	50%	70%	88%
Time trend (with or without by-clip variation)	42%	18%	25%	30%	75%
Either	75%	27%	63%	80%	100%
Possible convergence?	yes	no	yes	no	yes

ers' trajectories for different variables to influence each other, particularly for pairs where interaction is very frequent, or has particular social meaning. We might also expect trajectories to be expected by major events in the house that affect both patterns of interaction and social structure, such as housemate additions or evictions, or the introduction of an artificial social class distinction between housemates for one month (the heaven/hell split). No clear effects of either type have emerged so far, from qualitative inspection of the trajectories for groups of housemates who might be expected to influence each other (e.g., groups of housemates who grow particularly close). In future work we will look more carefully, using a detailed record of social interactions in the house.[13]

7.5.3 Dynamics of different variables

Examination of Fig. 7.9 suggests that variables show different patterns of time dependence. One way to compare variables is by the percentage of speakers who show change, in the sense of showing by-clip variation, time trends, or either. Another is whether convergence is observed or not. Both comparisons are given in Table 7.6.

With only five variables, we cannot say anything for certain about *why* variables show different patterns of time dependence. What we can do is evaluate several proposals from

[13] In a study of the dynamics of VOT for five housemates (Bane et al., 2010), we tentatively reported an effect of the heaven/hell split on VOT trajectories. However, the effect washes out in the current study, where more speakers are considered and by-clip variability is taken into account. This spurious result illustrates the importance of caution when proposing *causes* of observed patterns of time dependence, both in this and other longitudinal studies, given that many proposed explanations are post-hoc and based on relatively little data.

the short-term and long-term literatures, discussed in the previous chapter, for compatibility with our results.[14] Unfortunately, many other proposals (such as rule complexity) cannot be evaluated for our set of variables.

Social salience A role for social salience in which variables change has been advanced in both the short-term and long-term literatures. The standard argument is that more socially-salient variables are more likely to change, though the opposite view has also been proposed in some cases (see Sec. 4.3.1.3, 4.3.2.4). In our case, the most socially-salient variables are STRUT, and to a lesser extent TRAP′; CSD, VOT, and GOOSE are much less salient, if at all (see Sec. 4.4.3.1). And indeed, the percentage of speakers showing any change is higher for STRUT and TRAP′ than for CSD, VOT, or GOOSE. This pattern also holds for by-clip variation, though not for time trends.

Phonological status The phonological status of a variable is often invoked to explain differences in susceptibility to change, particularly in the long-term literature. As noted by Auer et al. (1998), what variables are phonological, and whether being phonological is a discrete property or a gradient scale, is highly theory-dependent. However, most authors would likely agree that variables which show only gradient, subphonemic variation are the 'least phonological'. This class certainly includes GOOSE and VOT: all speakers have the same number of categories, and differ only in their phonetic implementation. The same holds for TRAP′, although the lexical items corresponding to this vowel differ by speaker. Interestingly, these three variables are exactly those which show potential convergence. However, they are not more or less likely to change than CSD or STRUT, for any of the three types of change.

[14] The rationale for considering short-term explanations is that medium-term change could be the result of accumulated short-term changes. The rationale for considering long-term explanations is that whatever mechanism is responsible for some variables changing more than others over a timescale of years may apply over a timescale of three months as well.

Phonetic distance Phonetic distance, meaning some measure of perceptual or articulatory distinctness, is often invoked in both the short-term and long-term literatures: change is more likely for variables whose variants are further apart. This concept can be interpreted across multiple variables, or for different speakers within a single variable. At the level of all our variables, the prediction would be that more change should occur for variables with a broader range of variants. However, it is not clear how to compare ranges for different variables, since their distances are measured in different units (cf. Auer et al., 1998: 170).

Within a single variable, change is hypothesized to be more likely for speakers who are further away from the 'target'.[15] The cases of potential convergence we observed can interpreted this light. In each case, one or more housemates with the most extreme (highest or lowest) value for a variable showed a trend over the course of the season towards the variable's mean value in the house. Thus, it was the speakers with the greatest distance from the mean who changed towards it. However, in most cases of great phonetic distance, housemates did not show time trends towards the mean. The most striking case is STRUT, where there is a clear distinction between Northern English housemates, who use [ʊ] (Dale, Lisa, Luke), and other housemates, who use [ʌ] (Fig. 7.6). No speaker in either group shows a time trend towards the other group.

Contrast maintenance Recall that Nielsen (2011) showed in a short-term experiment that speakers would shift their VOTs for voiceless stops upwards (towards a model speaker with modified VOTs), but not downwards. Nielsen ascribed this asymmetry to contrast maintenance: decreasing VOT would endanger the contrast with voiced stops, while increasing VOT does not. If this effect persisted over many interactions, we would expect to observe a bias against decreasing VOT in medium-term trajectories. Instead, we see the opposite pattern (Fig. 7.2): three of the four speakers with time trends show a linear

[15] We assume the simplest possible interpretation of distance for each variable: the difference between two speakers in CSD rate, F2 for GOOSE, etc.

decrease in VOT, and the remaining speaker decreases following an increase (quadratic trend). Of course, there are many differences between our study and Nielsen's, beyond the timescale: the setting (laboratory vs. diary room), type of speech, American vs. British speakers. But the lack of a bias against decreasing VOT may suggest that contrast maintenance is not a concern over the medium term.

7.6 Discussion

The unique setting of the Big Brother house has allowed us to examine day-by-day trajectories of phonetic variables within individuals, for the first time. Our main finding is the existence of massive variability in these trajectories, over a period of three months; we described possible explanations for this variability, at the level of clips, housemates, and variables. The observed dynamics of variables in the house bear on a number of issues about short-term and long-term dynamics, and the relationship between them. Before turning to these, we discuss a few important caveats for interpreting our results.

7.6.1 Caveats

As a first study of medium-term time trajectories of linguistic variables, this study has necessarily been exploratory. We did not know what hypotheses to test prior to examining the data, and could only formulate explanations for patterns seen in the data post-hoc. As with any post-hoc explanation, the explanations for variability proposed in the previous section are only tentative. They must be tested on new datasets, or for different variables on the Big Brother data, and should only be generalized to other cases with caution. It is worth noting that this issue is not restricted to the current study, but applies to much work on longitudinal variation in individuals. Since we still know relatively little about change within individuals, many short-term and long-term studies are analyzed at least in part post-hoc; this issue is just not usually highlighted.

A second caveat relates to the robustness of our selection of a type of time dependence for each dynamic model. As noted above, the choice of model selection criterion matters; in particular, using BIC would give much sparser models than AIC. In ranking models by AIC, we found it was often the case that the second-best (and sometimes more) models differed relatively little in AIC from the top-ranked model. Both of these points suggest that we have sometimes not chosen a model of time dependence which is unequivocally better than others. Yet, we saw that our method generally predicts time dependence that matches the empirical data well, suggesting that when different models types of time dependence fit the data nearly as well as each other, it is because we had too little data to distinguish between them. To better characterize time dependence in future work, we could either use more data, or employ a model-averaging scheme, where different high-ranked models are combined (Burnham and Anderson, 2002). What impact would choosing the "wrong" time dependence have? At a coarse level, we may have sometimes concluded there was more change than there actually was, or concluded there was no change when in fact some occurred. Importantly, we saw empirically that our procedure for choosing a model of time dependence tends to be conservative: less change was predicted than was observed empirically. If anything, we may be underestimating the amount of change in some cases.

Finally, it is important to remember that all our conclusions are based on speech from the diary room, not interactions between housemates. We are using diary room speech to look for changes in housemates' baseline use of a variable. In doing so we abstract away from short-term shifts in social interaction, or from other sources, and ask whether we observe time dependence which might be due to an accumulation of short-term shifts. But our use of diary room speech means we cannot conclude anything about time dependence in how housemates talk *to each other*, a priori. In principle, a very different picture might emerge than for diary room speech. For example, clearer effects of social interaction might emerge if we looked just at conversations between a particular pair of housemates over time. Examining the dynamics of phonetic variables as used in social interactions in the house will be a primary focus of future work.

7.6.2 *The persistence hypothesis*

The persistence hypothesis is that short-term shifts that an individual makes during interaction can and do accumulate over time into long-term change. This hypothesis is implicitly assumed in much of the phonetic convergence literature, and explicitly assumed in the change-by-accommodation and identity projection models of the relationship between short-term and long-term change.[16] Stronger and weaker forms of the persistence hypothesis can be taken, depending on how widespread, accumulation is thought to be, across speakers and individuals. It is not entirely clear how strongly the hypothesis is meant in previous work. For example, Trudgill allowed room for variation in his earlier formulation, but more recently has emphasized the automaticity and pervasiveness of accommodation, which might be taken to suggest that accumulation is automatic as well.[17]

Under the strong form of the persistence hypothesis, we would expect to see pervasive change in the Big Brother house, given the extreme level of social interaction with the same people. This change could take the form of time trends or day-to-day variability, depending on how systematic the direction of short-term shift is in a particular housemate's use of a variable in interaction. But one way or another, there should be rampant time dependence. The observed dynamics do not support the strong persistence hypothesis, because we see *no* time dependence in a significant minority of cases. The most striking examples are trajectories for CSD: the 11 housemates considered have very different rates

[16] However, face-to-face interaction is not required in the identity projection model.

[17] "...accommodation may become permanent, particularly if attitudinal factors are favourable." (Trudgill, 1986: 39), versus "Linguistic diffusion and new-dialect formation are 'mechanical and inevitable' because linguistic accommodation is automatic, because, as Cappella (1997:69) says, it is an aspect of 'the relatively automatic behaviors manifested during social interaction'" (Trudgill, 2008: 252).

of usage (so there is ample room to change towards each other), but eight of them show no time dependence.

On the other hand, we do observe day-by-day variation and long-term trends in many cases, which *could* be due to accumulated short-term shifts. So our results are consistent with a weaker form of the persistence hypothesis, where speakers and variables show large differences in how much short-term shifts accumulate.

7.6.3 The stationarity hypothesis

The stationarity hypothesis is that individuals vary very little from day to day in their use of a variable, after controlling for conditioning factors such as speaking style. This hypothesis is necessary to conclude from long-term studies that a (statistically significant) difference observed in a speaker's use of a variable at two time points represents a meaningful change. The strong version of the hypothesis would be that individuals do not vary their baseline use of a variable at all from day to day. Assuming that at least some of the extensive by-clip variation we observe cannot be accounted for by style shifting, our results do not support the strong version. However, in a significant minority of cases no time dependence is observed, and by-clip variation occurs much more consistently across speakers for some variables (TRAP') than for others (CSD). So again, our results are consistent with a weaker version of the hypothesis: there is random day-to-day variation for only some speakers, for some variables.

The relevance of this finding depends on how much of the observed by-clip variation is truly due to day-by-day noise, and how much to style shifting. Controlled investigations of day-by-day variability are needed to establish its magnitude for different variables. The level of day-by-day variability is relevant not just for long-term studies, but the study of phonetic variability in general. If it turns out that significant day-by-day variability in phonetic variables is the norm, then previous work has likely overestimated the amount of by-speaker variability that exists in speech communities.

7.6.4 The relationship between short-term and long-term change

At the end of the previous chapter, we noted a tension between the literatures on short-term and long-term dynamics of phonetic variables. Short-term shifts are relatively robust across speakers and variables, while long-term shifts are highly irregular, with massive variability in outcomes for different speakers and variables. What is the relationship between the patterns seen in short-term and long-term dynamics?

Our approach to this question was to examine medium-term trajectories of phonetic variables. We found that speakers and variables show highly variable medium-term dynamics. A majority of cases were stable over the medium term, in the sense of showing no time dependence, or only by-clip variation. In a minority of cases, there was systematic change over time (time trends). These observations are strikingly similar to the results of long-term studies: massive variability among speakers and variables in how much change occurs, and relative stability in a majority of cases. The congruence between our results, on a timescale of months, and long-term studies, on a timescale of years, suggests that the disconnect between short-term and long-term dynamics is already present on a timescale shorter than months.

With this in mind, we can sketch a tentative account of the relationship between short-term and long-term dynamics in individuals, on several timescales:

1. *Interaction*: Some short-term shift during interaction nearly always occurs, across speakers and variables.

2. *Days*: A speaker's short-term shifts for a particular variable will often build up enough to affect her production from day to day, but whether this occurs varies by speaker and variable.

3. *Months–years*: Accumulation over the medium to long term occurs in a minority of cases, for some people and variables.

Why would the amount of accumulation differ by speaker and by variable, in (2) and (3)? Tentatively, some people have more plastic sound systems than others. Differences in plasticity could result from individual differences in mechanisms underlying perception and production, cognitive factors (e.g., aspects of a speaker's personality), or social factors (e.g., women might be more variable than men). Also tentatively, phonetic and phonological variables differ in how subject they are to medium-term change. The patterns we observed for different variables suggest two hypotheses. Social salience makes a variable more flexible from day-to-day, but not more likely to accumulate over months–years. Purely phonetic variables are more susceptible to actual convergence (change over the medium term towards a target) than "more phonological" variables.

More solid answers on why people and variables differ in plasticity must await future work. One intriguing possibility is that individual differences in the link between speech perception and production underly interspeaker differences in medium-term dynamics. Although short-term shifts occur robustly, individuals in short-term studies differ greatly in the *amount* of shift. Perhaps the speakers who change less over the medium term are precisely those who shift less during interaction. Alternatively, perhaps speakers differ in how much short-term shifts persist from one interaction to the next. (In exemplar-theoretic terms, speakers may vary in the decay rate for exemplars encountered in perception.) The first possibility can be tested in the Big Brother data by examining the size of short-term shifts different housemates show for our five variables during interactions, and comparing to their medium-term patterns; this is a primary direction for future work. The second possibility cannot be tested for Big Brother, since in general we only have access to isolated conversations (rather than a continuous sequence).

The account sketched in (1)–(3) is admittedly very general, even vague. We believe that much more empirical work documenting longitudinal variation in individuals is needed before making stronger hypotheses about its sources, because the basic empirical facts are far from clear at this point. Much of the existing literature assumes that a relatively straightforward link between short-term and long-term change exists, and certainly that some link *does* exist. Yet the current study suggests that the dynamics of phonetic variables over months, even in a setting of extreme social interaction, are highly complex. In our opinion, current accounts of the link between short-term and long-term change may be overly ambitious, given how little is currently known about longitudinal variation in individuals.

Our tentative tone should not be taken to mean that we believe short-term shifts are not linked to long-term change, but rather that assuming there is a link, it is indirect and poorly understood so far. Although the current study considers only adults, these points by and large hold for adolescents as well.[18] In general, a vast amount remains to be learned about phonetic and phonological dynamics in individuals at different timescales, and their sources. Thanks to previous work, a good deal is known about dynamics on very short and very long timescales. To understand the relationship between them, it is necessary to examine how speakers change on timescales in between. This project has been one step in this direction.

[18] Longitudinal studies of children are much more common. Particularly notable is a remarkable project by Wolfram and colleagues which followed African American Vernacular English speakers from ages 1–17 (van Hofwegen and Wolfram, 2010).

CHAPTER 8

CONCLUSION

This thesis has addressed phonetic and phonological dynamics in individuals during adulthood, as well as the related areas of automatic phonetic measurement, and synchronic variation in spontaneous speech. Our main contributions fall into three areas, and suggest directions for future work in each.

Medium-term phonetic and phonological dynamics The most important contribution of this thesis is a medium-term study of phonetic and phonological dynamics in individuals, over a period of three months. This study is the first (to our knowledge) to examine day-by-day variation in the sound systems of individuals. Our results suggest a possible link between the different dynamics observed in short-term and long-term studies, rooted in huge differences among speakers and variables in how much and what type of change occurs in the medium term.

However, our study perhaps raises more questions about longitudinal variation than it answers. We allowed for four qualitatively different patterns of time dependence for variables within individual speakers—no change, by-clip variation, time trends, and both—and found many cases of each. The differences among speakers in characteristic patterns of time dependence are particularly intriguing. Their causes might lie in patterns of social interaction in the house, social factors (e.g., the social roles individuals come to play in the house), or individuals might differ in the plasticity of their sound systems. Exploring these different options is an important direction for future work, both on the Big Brother corpus and in other datasets.

The existence of time trends in our data shows that short-term shifts during conversation for some speakers *could* have accumulated into change over weeks, but not whether they did. An important direction for future work will be analyzing whether speakers who show more change in the medium term are also those who change more in short-term interactions with other housemates; if true, this would provide more convincing (though still indirect) evidence for the accumulation of short-term shifts within housemates. More generally, experimental studies are needed to directly test whether short-term shifts ever accumulate, and on what timescale they persist; to my knowledge, there are no studies addressing the persistence of short-term shifts in phonetic or phonological variables for more than an hour. This is an important direction for future work, given that persistence of short-term shifts is (tacitly or explicitly) assumed by many scholars.

In many cases, speakers on Big Brother did change their use of a variable over three months. An important question raised by our results is how deeply such change affects speakers' sound systems. For each variable, we tested for time dependence in each speaker's baseline value, after controlling for the effects of conditioning factors. But we do not know whether the effects of the conditioning factors *themselves* changed over time. Intuitively, how a variable is conditioned constitutes a deeper part of a speaker's linguistic knowledge than its baseline value, and is more stable across situations and speakers: the qualitative effects of conditioning factors (i.e., whether a following consonant promotes or inhibits CSD) for a given variable tend to not vary for the same speaker across different speaking styles, or across speakers in a speech community, while the baseline value does.[1] Analogously, perhaps baseline values are much more prone to longitudinal variation than conditioning factors. An important direction for future work is testing whether the probabilistic grammar characterizing a speaker's use of a variable can change over time. Testing this hypothesis in spontaneous speech will require huge datasets, because of the sparsity of many conditioning factors. (For example, the rarity of past tenses relative to monomorphemes would make studying the dynamics of the effect of morphology on CSD very difficult.) Experimental studies, where the effect of a conditioning factor

[1] c.f. Lim and Guy (2005: 157): "The linguistic unity of speech communities lies in shared linguistic practices and evaluations. Where variable processes are concerned, this linguistic unity extends to shared constraint effects.... Although rarely explicitly stated, the conventional practice in sociolinguistic research is to assume that linguistic constraint effects are stable across different speech styles."

can be fully assessed at several points in time, can more directly address how plastic individuals are in this respect, but with less ecological validity.

Using natural experiments to study dynamics Another contribution of this thesis is to demonstrate the utility of natural experiments for studying variation and change. In particular, a season of Big Brother served as an excellent laboratory for studying medium-term change in sound systems. The Big Brother franchise holds great potential for studying the dynamics of linguistic systems more generally. Unlike many natural experiments, Big Brother is quasi-replicable: there have now been hundreds of seasons of shows in many languages, with relatively similar formats. Thus, a surprising finding or a tentative hypothesis based on data from one season can be tested on data from another season.

Natural experiments such as Big Brother are particularly valuable for examining *trajectories* of linguistic variables across many timepoints. The day-by-day resolution of Big Brother is exceptional, and probably not replicable in any real-world setting. But many other natural experiments allow the possibility of observing trajectories of linguistic variables at a coarser resolution, such as university campuses, high schools, study abroad programs, or internet forums; recent studies have begun to explore trajectories of linguistic variables in such settings (e.g., Altmann et al., 2009; Chang, 2012; Orr et al., 2011). Why examine trajectories? One motivation is empirical: we simply know very little about what trajectories of linguistic variables look like, especially for individual speakers. Filling this lacuna matters because what happens between the endpoints is crucial for answering fundamental questions about variation and change, such as those addressed here: the relationship between short-term and long-term dynamics, to what extent short-term shifts accumulate, and how much day-by-day fluctuation individuals show in their use of linguistic variables.

The promise and challenges of automatic phonetic measurement The main methodological contribution of this thesis is a new tool for automatic VOT measurement. Our approach combines knowledge about the cues human annotators use to measure VOT with machine learning techniques for predicting structured output, to tailor a VOT measurement algorithm that meets the three criteria laid out in the introduction: accuracy, trainability, and robustness. Combining knowledge about the annotation task to be performed with appropriate machine learning techniques is a promising direction for designing algorithms to automate phonetic measurement in general.

Such tools are needed to scale up studies in phonetics and sociolinguistics. Mark Liberman has recently argued that thanks to the advent of large speech corpora and the computational methods to examine them, we are on the verge of a "golden age of speech and language science", where linguists will be able to ask fundamentally new questions about speech production, and language more generally.[2] In Liberman's view, the type of large-scale phonetic study made possible by new corpora and methods consists of four steps: forced alignment, pronunciation modeling, automatic phonetic measurement, and mixed-effects regression models.

However, the third step is highly non-trivial. Though off-the-shelf algorithms can be used to measure some aspects of speech (especially suprasegmental variables, such as pitch), in general they will not suffice to replace the kind of skilled manual measurements, especially of subphonemic quantities, which underlie speech production research. To realize Liberman's vision, speech scientists will need to create or adapt methods from the ASR and machine learning literatures, and extensively test them against human measurements. In general, phonetic measurement is a fertile and largely unexplored application area for ASR and machine learning.

[2] For example, "The Golden Age of Speech and Language Science", at `http://languagelog.ldc.upenn.edu/myl/Groningen.pdf`.

REFERENCES

Abramowicz, Ł. (2007). Sociolinguistics meets exemplar theory: Frequency and recency effects in (ing). *University of Pennsylvania Working Papers in Linguistics*, 13(2):27–37.

Abrego-Collier, C., Grove, J., Sonderegger, M., and Yu, A. (2011). Effects of speaker evaluation on phonetic convergence. In Lee and Zee (2011), pages 192–195.

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(5):3099–3107.

Adjarian, H. (1899). Les explosives de l'ancien arménien étudiées dans les dialectes modernes. *La parole. Revue internationale de Rhinologie, Otologie, Laryngologie et Phonétique expérimentale*, pages 119–127.

Ali, A. (1999). *Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition*. PhD thesis, University of Pennsylvania.

Allen, J., Miller, J., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustic Society of America*, 113(1):544–552.

Altmann, E., Pierrehumbert, J., and Motter, A. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One*, 4(11):e7678.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

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

Auer, P. (2005). Syntax als prozess. *InLiSt - Interaction and Linguistic Structures*, 41.

Auer, P. (2010). Why are increments such elusive objects? an afterthought. *Pragmatics*, 17(4):647–658.

Auer, P., Barden, B., and Großkopf, B. (1998). Subjective and objective parameters determining 'salience' in long-term dialect accommodation. *Journal of Sociolinguistics*, 2(2):163–187.

Auer, P. and Hinskens, F. (2005). The role of interpersonal accommodation in a theory of language change. In Auer, P., Hinskens, F., and Kerswill, P., editors, *Dialect change: convergence and divergence in European languages*, pages 335–357. Cambridge University Press, Cambridge.

Auzou, P., Ozsancak, C., Morris, R., Jan, M., Eustache, F., and Hannequin, D. (2000). Voice onset time in aphasia, apraxia of speech and dysarthria: a review. *Clinical Linguistics & Phonetics*, 14(2):131–150.

Baayen, R. (2008). *Analyzing linguistic data*. Cambridge University Press, Cambridge.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Baayen, R., Piepenbrock, R., and Gulikers, L. (1996). *CELEX2*. Linguistic Data Consortium, Philadelphia.

Babel, M. (2009). *Phonetic and Social Selectivity in Speech Accommodation*. PhD thesis, University of California at Berkeley.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, 39(4):437–456.

Babel, M. (2011). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1):177–189.

Bailey, G. and Tillery, J. (1999). The Rutledge Effect: The impact of interviewers on survey results in linguistics. *American Speech*, 74:389–402.

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

Bane, M., Graff, P., and Sonderegger, M. (2010). Longitudinal phonetic variation in a closed system. *Proceedings of the Annual Meeting of the Chicago Linguistics Society*, 46. In press.

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.

Barden, B. and Großkopf, B. (1998). *Sprachliche Akkommodation und soziale Integration [Speech Accommodation and Social Integration]*. Niemeyer, Tübingen.

Bargh, J. and Chartrand, T. (1999). The unbearable automaticity of being. *American Psychologist*, 54(7):462–479.

Bates, D. (2011). Linear mixed model implementation in lme4. Manuscript, University of Wisconsin.

Bates, D., Maechler, M., and Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-40.

Bauer, L. (1985). Tracing phonetic change in the received pronunciation of British English. *Journal of Phonetics*, 13(1):61–81.

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

Beal, J. (2004). English dialects in the North of England: phonology. In Kortmann and Schneider (2004), pages 113–133.

Beckman, M. (1997). A typology of spontaneous speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing prosody: Computational models for processing spontaneous speech*, pages 7–26. Springer-Verlag, New York.

Beddor, P. (1982). *Phonological and phonetic effects of nasalization on vowel height*. PhD thesis, University of Minnesota. Reproduced by Indiana University Linguistics Club.

Beddor, P., Harnsberger, J., and Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics*, 30(4):591–627.

Bell, A. (2001). Back in style: Reworking audience design. In Eckert, P. and Rickford, J., editors, *Style and Sociolinguistic Variation*, pages 139–169. Cambridge University Press, Cambridge.

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

Bergmann, P. (2006). Regional variation in intonation. In *Language variation–European perspectives: selected papers from the Third International Conference on Language Variation in Europe (ICLaVE 3), Amsterdam, June 2005*, pages 23–36, Amsterdam. John Benjamins.

Blondeau, H. (2001). Real-time changes in the paradigm of personal pronouns in Montreal French. *Journal of Sociolinguistics*, 5(4):453–474.

Blondeau, H. (2006). La trajectoire de l'emploi du futur chez une cohorte de Montréalais francophones entre 1971 et 1995. *Canadian Journal of Applied Linguistics*, 37(2):73–98.

Blondeau, H., Sankoff, G., and Charity, A. (2002). Parcours individuels dans deux changements linguistiques en cours en français montréalais. *Revue québécoise de linguistique*, 31(1):13–38.

Bloomfield, L. (1933). *Language.* Henry Holt, New York.

Boersma, P. and Weenink, D. (2011). Praat: doing phonetics by computer (Version 5.2.17) [Computer program].

Bowie, D. (2005). Language change over the lifespan: A test of the apparent time construct. *University of Pennsylvania Working Papers in Linguistics*, 11(2):45–58.

Braun, A. (1983). VOT im 19. Jahrhundert oder "Die Wiederkehr des Gleichen". *Phonetica*, 40(4):323–327.

Brink, L. and Lund, J. (1975). *Dansk rigsmål: lydudviklingen siden 1840 med særligt henblik på sociolekterne i København*. Gyldendal, København.

British Library (2007). The trap∼bath split. [Online; accessed 10-January-2012; created 13-March-2007].

Brouwer, S., Mitterer, H., and Huettig, F. (2010). Shadowing reduced speech and alignment. *Journal of the Acoustical Society of America*, 128(1):EL32–EL37.

Brugnara, F., Falavigna, D., and Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–370.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

Burnham, K. and Anderson, D. (2004). Multimodel Inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.

Bybee, J. (2000). The phonology of the lexicon: evidence from lexical diffusion. In Barlow, M. and Kemmer, S., editors, *Usage-based models of language*, pages 65–85. CSLI, Stanford.

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford University Press, New York.

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

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

Caramazza, A. and Yeni-Komshian, G. (1974). Voice onset time in two French dialects. *Journal of Phonetics*, 2:239–245.

Chambers, J. (1992). Dialect acquisition. *Language*, 68(4):673–705.

Chambers, J. (2003). *Sociolinguistic theory: Linguistic variation and its social significance*. Wiley-Blackwell, Malden, MA.

Chang, C. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics*, 40(2).

Chartrand, T. and Bargh, J. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

Chistovich, L., Fant, G., de Serpa-Leitão, A., and Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm*, 2:1–18.

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

Clark, L. and Trousdale, G. (2009). Exploring the role of token frequency in phonological change: evidence from TH-Fronting in east-central Scotland. *English Language and Linguistics*, 13(1):33–55.

Clark, U. (2004). The English West Midlands: phonology. In Kortmann and Schneider (2004), pages 134–162.

Coetzee, A. and Pater, J. (2011). The place of variation in phonological theory. In Goldsmith et al. (2011), pages 401–434.

Cohen, J., Cohen, P., West, S., and Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.

Cole, J. and Hasegawa-Johnson, M. (2012). Corpus phonology with speech resources. In Cohn, A., Fougeron, C., and Huffman, M., editors, *The Oxford Handbook of Laboratory Phonology*, pages 431–440. Oxford University Press, Oxford.

Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, 38(2):167–184.

Cole, J. and Shattuck-Hufnagel, S. (2011). The phonology and phonetics of perceived prosody: What do listeners imitate? In *Proceedings of INTERSPEECH-2011*, pages 969–972.

Cooper, A. (1991). *An articulatory account of aspiration in English*. PhD thesis, Yale University.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language*, 46:49–70.

Cox, F. and Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association*, 37(3):341–350.

Cramer, J. (2010). *The effect of borders on the linguistic production and perception of regional identity in Louisville, Kentucky*. PhD thesis, University of Illinois at Urbana-Champaign.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Crystal, T. and House, A. (1988a). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, 83(4):1553–1573.

Crystal, T. and House, A. (1988b). The duration of American-English stop consonants: an overview. *Journal of Phonetics*, 16:285–294.

Daleszynska, A. (2011). Whats gender got to do with it? Investigating the effect of gender, age and place on/t, d/deletion in Bequia. In *Proceedings of the Summer School of Sociolinguistics*. University of Edinburgh, Edinburgh.

Daveluy, M. (1988). L'usage des déterminants démonstratifs dans la communauté francophone de Montréal en 1971 et en 1984. Master's thesis, Université de Montréal.

Dekel, O., Keshet, J., and Singer, Y. (2004). Large margin hierarchical classification. In *Proceedings of the 21st International Conference on Machine Learning*, pages 209–216.

Delvaux, V. and Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, 64(2-3):145–173.

Dijksterhuis, A. and Bargh, J. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, 33:1–40.

Dilley, L. and Pitt, M. (2007). A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *Journal of the Acoustical Society of America*, 122(4):2340–2353.

Dinkin, A. (2008). The real effect of word frequency on phonetic variation. *University of Pennsylvania Working Papers in Linguistics*, 14(1):97–106.

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

Docherty, G., Watt, D., Llamas, C., Hall, D., and Nycz, J. (2011). Variation in voice onset time along the Scottish-English border. In Lee and Zee (2011), pages 591–594.

Dubois, S. (1992). Extension particles, etc. *Language Variation and Change*, 4(2):163–203.

Ernestus, M. and Baayen, R. (2011). Corpora and exemplars in phonology. In Goldsmith et al. (2011), pages 374–400.

Ernestus, M., Lahey, M., Verhees, F., and Baayen, R. (2006). Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America*, 120(2):1040–1051.

Evanini, K., Isard, S., and Liberman, M. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. In *Proceedings of INTERSPEECH-2009*, pages 1655–1658.

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

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

Fasold, R. (1972). *Tense marking in Black English: A linguistic and social analysis*. Center for Applied Linguistics, Washington, DC.

Ferragne, E. and Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, 40(1):1–34.

Fidelholtz, J. (1975). Word frequency and vowel reduction in English. In *Chicago Linguistic Society*, volume 11, pages 200–213.

Foreman, A. (2003). *Pretending to be someone youre not: A study of second dialect acquisition in Australia*. PhD thesis, Monash University.

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

Fowler, C., Brown, J., Sabadini, L., and Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, 49(3):396–413.

Fowler, C., Sramko, V., Ostry, D., Rowland, S., and Hallé, P. (2008). Cross language phonetic influences on the speech of French-English bilinguals. *Journal of Phonetics*, 36(4):649–663.

Francis, A., Ciocca, V., and Yu, J. (2003). Accuracy and variability of acoustic measures of voicing onset. *Journal of the Acoustic Society of America*, 113(2):1025–1032.

Fridland, V. (2008). Patterns of /uw/, /ʊ/, and /ow/ fronting in Reno, Nevada. *American Speech*, 83(4):432–454.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, 63(1):223–230.

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

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, J., and Coupland, N., editors, *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1–68. Cambridge Univ Press, Cambridge.

Giles, H. and Powesland, P. (1975). Speech style and social evaluation.

Gilles, P. (1999). *Dialektausgleich im Lëtzebuergeschen [Dialect Shift in Luxembourgish]*. Niemeyer, Tübingen.

Godfrey, J. and Holliman, E. (1997). *Switchboard-1 Release 2*. Linguistic Data Consortium, Philadelphia.

Goeman, A. (1999). *T-deletie in Nederlanse dialecten: kwantitatieve analyse van structurele, ruimtelijke en temporele variatie [T-deletion in Dutch dialects. Quantitative analysis of structural geographical and temporal variation]*. Holland Academic Graphics, The Hague.

Goldinger, S. (2000). The role of perceptual episodes in lexical processing. In *Proceedings of the Workshop on Spoken Word Access Processes*, pages 155–158, Nijmegen. Max Planck Institute.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105:251–279.

Goldsmith, J., Riggle, J., and Yu, A., editors (2011). *The Handbook of Phonological Theory*. Blackwell, 2nd edition.

Gordon, C. (2011). Impression management on reality TV: Emotion in parental accounts. *Journal of Pragmatics*, pages 3551–3564.

Gregersen, F., Maegaard, M., and Pharao, N. (2009). The long and short of (æ)-variation in Danish–a panel study of short (æ)-variants in Danish in real time. *Acta Linguistica Hafniensia*, 41(1):64–82.

Guy, G. (1980). Variation in the group and the individual: the case of final stop deletion. In Labov (1980), pages 1–36.

Guy, G. (1988). Advanced VARBRUL analysis. In Ferrara, K., Brown, B., Walters, K., and Baugh, J., editors, *Linguistic change and contact*, pages 124–136. Department of Linguistics, University of Texas at Austin, Austin, TX.

Guy, G. (1991). Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change*, 3(1):1–22.

Halberstadt, A. (1998). *Heterogeneous measurements and multiple classifiers for speech recognition*. PhD thesis, Massachusetts Institute of Technology.

Hale, M. and Reiss, C. (2000). Phonology as cognition. In Burton-Roberts, N., Carr, P., and Docherty, G., editors, *Phonological Knowledge: Conceptual and Empirical Numbers*, pages 161–184. Oxford University Press, Oxford.

Hansen, J., Gray, S., and Kim, W. (2010). Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification. *Speech Communication*, 52(10):777–789.

Harrell, F. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Verlag, New York.

Harrington, J. (2006). An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics*, 34(4):439–457.

Harrington, J. (2007). Evidence for a relationship between synchronic variability and diachronic change in the Queens annual Christmas broadcasts. In Cole, J. and Hualde, J., editors, *Laboratory Phonology 9*, pages 125–144. Mouton de Gruyter.

Harrington, J., Palethorpe, S., and Watson, C. (2000a). Does the Queen speak the Queen's English? *Nature*, 408(6815):927–928.

Harrington, J., Palethorpe, S., and Watson, C. (2000b). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association*, 30(1-2):63–78.

Harrington, J., Palethorpe, S., and Watson, C. (2005). Deepening or lessening the divide between diphthongs? An analysis of the Queen's annual Christmas broadcasts. In *A figure of speech: A Festschrift for John Laver*, pages 227–261. Lawrence Erlbaum, Mahwah, NJ.

Harrington, J., Palethorpe, S., and Watson, C. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *Proceedings of INTERSPEECH-2007*, pages 2753–2756.

Hawkins, S. and Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, 35(2):183–199.

Hay, J., Jannedy, S., and Mendoza-Denton, N. (1999). Oprah and /ay/: Lexical frequency, referee design, and style. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 1389–1392.

Hay, J. and Sudbury, A. (2005). How rhoticity became /r/-sandhi. *Language*, 81(4):799–823.

Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34(4):458–484.

Hazen, K. (2011). Flying high above the social radar: Coronal stop deletion in modern Appalachia. *Language Variation and Change*, 23(1):105–137.

Heselwood, B. and McChrystal, L. (2000). Gender, accent features and voicing in Panjabi-English bilingual children. *Leeds Working Papers in Linguistics and Phonetics*, 8:45–70.

Hillenbrand, J., Clark, M., and Nearey, T. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109(2):748–763.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111.

Hinskens, F. (1996). *Dialect levelling in Limburg: Structural and sociolinguistic aspects*. Niemeyer, Tübingen.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. Routledge, New York.

Jacewicz, E., Fox, R., O'Neill, C., and Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2):233–256.

Johnson, D. E. (2012). Progress in regression: why sociolinguistic data calls for mixed models. Submitted Ms.

Johnson, K. (2011). *Quantitative methods in linguistics*. Wiley-Blackwell, Malden, MA.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 229–254. John Benjamins, Amsterdam.

Kappes, J., Baumgaertner, A., Peschke, C., and Ziegler, W. (2009). Unintended imitation in nonword repetition. *Brain and language*, 111(3):140–151.

Kazemzadeh, A., Tepperman, J., Silva, J., You, H., Lee, S., Alwan, A., and Narayanan, S. (2006). Automatic detection of voice onset time contrasts for use in pronunciation assessment. In *Proceedings of INTERSPEECH-2006*, pages 721–724.

Keating, P. (1998). Word-level phonetic variation in large speech corpora. In Alexiadou, A., N.Fuhrop, Kleinhenz, U., and Law, P., editors, *ZAS Papers in Linguistics 11*, pages 35–50. ZAS, Berlin.

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(2):131–142.

Kendall, T. (2011). Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada*, 11(2):361–389.

Kerswill, P. and Williams, A. (2002). "Salience" as an explanatory factor in language change: evidence from dialect levelling in urban England. In Jones, M. and Esch, E., editors, *Language Change: The Interplay of Internal, External and Extra-Linguistic Factors*, pages 81–110. Mouton de Gruyter, Berlin.

Keshet, J., Shalev-Shwartz, S., Singer, Y., and Chazan, D. (2007). A large margin algorithm for speech-to-phoneme and music-to-score alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2373–2382.

Kessinger, R. and Blumstein, S. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, 25(2):143–168.

Keune, K., Ernestus, M., Van Hout, R., and Baayen, R. (2005). Variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory*, 1(2):183–223.

Khan, F. (1991). Final consonant cluster simplification in a variety of Indian English. In Chesire, J., editor, *English around the world: Sociolinguistic perspectives*, pages 288–98. Cambridge University Press, Cambridge.

Kim, M. (2011). Phonetic convergence after perceptual exposure to native and nonnative speech: Preliminary findings based on fine-grained acoustic-phonetic measurement. In Lee and Zee (2011), pages 1074–1077.

Kim, M., Horton, W., and Bradlow, A. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, 2(1):125–156.

Kingston, J. (2007). The phonetics-phonology interface. In de Lacy, P., editor, *The Cambridge Handbook of Phonology*, pages 435–456. Cambridge University Press, Cambridge.

Klatt, D. (1973). Durational characteristics of prestressed word-initial consonant clusters in English. *MIT Research Laboratory of Electronics Quartelry Progress Report*, 108:253–260.

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

Kortmann, B. and Schneider, E. (2004). *A Handbook of Varieties of English*, volume 1: Phonology. de Gruyter, New York.

Kramer, M. (2005). R^2 statistics for mixed models. In *Proceedings of the Conference on Applied Statistics in Agriculture*, volume 17, pages 148–160.

Kurki, T. (2003). Applying the apparent-time method and the real-time method on Finnish. In *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe. ICLaVE 2*, pages 241–252. Dept. of Scandinavian Languages, Uppsala University, Uppsala.

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.

Labov, W. (1972). Some principles of linguistic methodology. *Language in Society*, 1(1):97–120.

Labov, W., editor (1980). *Locating Language in Time and Space*. Academic Press, New York.

Labov, W. (1994). *Principles of linguistic change. Volume 1: Internal factors*. Blackwell, Oxford.

Labov, W. (2000). *Principles of linguistic change. Volume 2: Social factors*. Blackwell, Oxford.

Labov, W. (2001). Applying our knowledge of African American English to the problem of raising reading levels in inner city schools. In Lanehart, S., editor, *Sociocultural and historical contexts of African American English*, pages 299–330. John Benjamins, Philadelphia.

Labov, W. (2010). *Principles of linguistic change. Volume 3: Cognitive and cultural factors*. Blackwell, Oxford.

Labov, W., Ash, S., and Boberg, C. (2006). *The atlas of North American English: phonetics, phonology, and sound change*. Mouton de Gruyter.

Labov, W. and Auger, J. (1993). The effect of normal aging on discourse: A sociolinguistic approach. In Joanette, H. B. . Y., editor, *Discourse in neurologically impaired and normal aging adults*, pages 115–133. Singular, San Diego, CA.

Labov, W. and Cohen, P. (1967). Systematic relations of standard and non-standard rules in the grammar of Negro speakers. In *Project Literacy Reports No. 8*, pages 66–84. Cornell University, Ithaca, NY.

Labov, W., Cohen, P., and Robins, C. (1965). *A preliminary study of the structure of English used by Negro and Puerto Rican speakers in New York City*. Final report, Cooperative Research Project 3091. [ERIC ED003819].

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*. U.S. Regional Survey, Philadelphia. Final report, Cooperative Research Project 3288. 2 vols.

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

Lawson, E., Scobbie, J., and Stuart-Smith, J. (2011). A single case study of articulatory adaptation during acoustic mimicry. In Lee and Zee (2011), pages 1170–1173.

Le Page, R. and Tabouret-Keller, A. (1985). *Acts of identity: Creole-based approaches to language and ethnicity*. Cambridge University Press, Cambridge.

Lee, W.-S. and Zee, E., editors (2011). *Proceedings of the 17th International Congress of Phonetic Sciences*.

Lelong, A. and Bailly, G. (2011). Study of the phenomenon of phonetic convergence thanks to speech dominoes. (6800):273–286.

Lelong, A. and Bailly, G. (2012). Characterizing phonetic convergence with speaker recognition techniques. In *Proceedings of the Listening Talker Workshop*, pages 28–31, Edinburgh.

Lessard, P. (1989). Variabilité linguistique et variabilité sociale dans la communauté francophone de Montréal. Master's thesis, Université de Montréal.

Levon, E. (2006). Mosaic identity and style: Phonological variation among Reform American Jews. *Journal of Sociolinguistics*, 10(2):181–204.

Levy, R. (2012). *Probabilistic methods in the study of langauge*. May 17, 2012 version of draft in progress.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence*. PhD thesis, Universität Stuttgart.

Lewandowski, N. and Dogil, G. (2010). Identity negotiation in native-nonnative dialogs: Quantifying phonetic adaptation. In De Cillia, R., Gruber, H., Krzyzanowski, M., and Menz, F., editors, *Diskurs, Politik, Identität: Festschrift für Ruth Wodak*, pages 389–399. Mouton de Gruyter, Berlin.

Lim, L. and Guy, G. (2005). The limits of linguistic community: speech styles and variable constraint effects. *University of Pennsylvania Working Papers in Linguistics*, 10:157–170.

Lin, C. and Wang, H. (2011). Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection. *Journal of the Acoustical Society of America*, 130(1):514–525.

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

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

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

Lorenzo-Dus, N. (2009). "You're barking mad, I'm out": Impoliteness and broadcast talk. *Journal of Politeness Research. Language, Behaviour, Culture*, 5(2):159–187.

MacCallum, R., Zhang, S., Preacher, K., and Rucker, D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1):19.

MacKenzie, L. and Sankoff, G. (2010). A quantitative analysis of diphthongization in Montreal French. *University of Pennsylvania Working Papers in Linguistics*, 15(2):91–100.

Magee, L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests. *American Statistician*, 44(3):250–253.

Maindonald, J. and Braun, J. (2007). *Data analysis and graphics using R: an example-based approach*. Cambridge Univ Press, Cambridge.

Masuya, Y. (1997). Voice Onset Time of the Syllable-Initial /p/, /t/ and /k/ Followed by an Accented Vowel in Lowland Scottish English. In *Onseigaku to oninron: shuyo ronko* [Phonetics and phonology: selected papers], pages 139–172. Kobian Shobo, Tokyo.

McCrea, C. R. and Morris, R. J. (2005). The effects of fundamental frequency level on voice onset time in normal adult male speakers. *Journal of Speech, Language and Hearing Research*, 48:1013–1024.

Meyerhoff, M. (1998). Accommodating your data: The use and misuse of accommodation theory in sociolinguistics. *Language & Communication*, 18(3):205–25.

Michelas, A. and Nguyen, N. (2011). Uncovering the effect of imitation on tonal patterns of French Accentual Phrases. In *Proceedings of INTERSPEECH-2011*, pages 973–976.

Miller, J., Green, K., and Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, 43(1-3):106–115.

Mitterer, H. and Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, 109(1):168–173.

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

Munro, M., Derwing, T., and Flege, J. (1999). Canadians in Alabama: A perceptual study of dialect acquisition in adults. *Journal of Phonetics*, 27(4):385–403.

Myers, J. and Guy, G. (1997). Frequency effects in variable lexical phonology. *University of Pennsylvania Working Papers in Linguistics*, 4(1):215–27.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nahkola, K. and Saanilahti, M. (2004). Mapping language changes in real time: A panel study on Finnish. *Language Variation and Change*, 16(2):75–92.

Namy, L., Nygaard, L., and Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21(4):422–432.

Naro, A. and Scherre, M. (2003). Estabilidade e mudança lingüística em tempo real: a concordância de número. In De Pavia, M. & Duarte, M., editor, *Mudança lingüística em tempo real*, pages 47–62. Capa, Rio de Janeiro.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Nearey, T. and Rochet, B. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association*, 24(1):1–19.

Neu, H. (1980). Ranking of constraints on /t,d/ deletion in American English: A statistical analysis. In Labov (1980), pages 37–54.

Nielsen, K. (2008). *Word-level and Feature-level Effects in Phonetic Imitation*. PhD thesis, University of California at Los Angeles.

Nielsen, K. (2010). Phonetic imitation of Japanese vowel devoicing. In *Proceedings of INTERSPEECH-2010*, pages 1553–1556.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39(2):132–142.

Nilsenová, M., Swerts, M., Houtepen, V., and Dittrich, H. (2009). Pitch adaptation in different age groups: boundary tones versus global pitch. In *Proceedings of INTERSPEECH-2009*, pages 1015–1018.

Nilsenová, M. and van Amelsvoort, M. (2010). Syntax drives phonological choice–even independently of word choice. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 212–217, Austin, TX. Cognitive Science Society.

Niyogi, P. and Ramesh, P. (1998). Incorporating voice onset time to improve letter recognition accuracies. In *Proceedings of ICASSP-98*, pages 13–16.

Niyogi, P. and Ramesh, P. (2003). The voicing feature for stop consonants: Recognition experiments with continuously spoken alphabets. *Speech Communication*, 41(2-3):349–367.

Ohala, J. (1981). Articulatory constraints on the cognitive representation of speech. In Myers, T., Laver, J., and Anderson, J., editors, *The Cognitive Representation of Speech*, pages 111–122. North Holland, New York.

Ohala, J. (1990). There is no interface between phonology and phonetics: a personal view. *Journal of Phonetics*, 18(2):153–172.

Oostdijk, N. (2000). The spoken Dutch corpus: overview and first evaluation. In *Proceedings of LREC-2000*, volume 2, pages 887–894.

Orr, R., Quené, H., Beek, R., Diefenbach, T., Leeuwen, D., and Huijbregts, M. (2011). An international English speech corpus for longitudinal study of accent development. In *Proceedings of INTERSPEECH-2011*.

Pak, M., editor (2009). *Current Numbers in Unity and Diversity of Languages*. Linguistic Society of Korea, Seoul.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119(4):2382–2393.

Pardo, J. (2009). Expressing oneself in conversational interaction. In Morsella, E., editor, *Expressing Oneself/Expressing Ones self: Communication, Cognition, Language, and Identity*, pages 183–196. Lawrence Erlbaum, Mahwah, NJ.

Pardo, J., Gibbons, R., Suppes, A., and Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, 40(1):190–197.

Pardo, J., Jay, I., and Krauss, R. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, 72(8):2254–2264.

Paterson, N. (2011). *Interactions in Bilingual Speech Processing*. PhD thesis, Northwestern University.

Patrick, P. (1991). Creoles at the intersection of variable processes: (TD)-deletion and past-marking in the Jamaican mesolect. *Language Variation and Change*, 3(2):171–189.

Paul, H. (1880). *Prinzipien der Sprachgeschichte*. Max Niemeyer, Halle, 1st edition.

Payne, A. (1976). *The acquisition of the phonological system of a second dialect*. PhD thesis, University of Pennsylvania.

Payne, A. (1980). Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In Labov (1980), pages 143–178.

Penhallurick, R. (2004). Welsh English: phonology. In Kortmann and Schneider (2004), pages 98–112.

Peterson, G. and Barney, H. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.

Picheny, M., Durlach, N., and Braida, L. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29(4):434–446.

Pickering, M. and Ferreira, V. (2008). Structural priming: a critical review. *Psychological Bulletin*, 134(3):427–459.

Pickering, M. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Pierrehumbert, J. (2002). Word-specific phonetics. In Gussenhoven, C. and Warner, N., editors, *Laboratory Phonology 7*, pages 101–139. Mouton de Gruyter.

Pinheiro, J. and Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer, New York.

Pisoni, D. (1980). Variability of vowel formant frequencies and the quantal theory of speech: a first report. *Phonetica*, 37(5-6):285–305.

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Department of Psychology, Ohio State University, Columbus.

Podesva, R. (2006). *Phonetic Detail in Sociolinguistic Variation*. PhD thesis, Stanford University.

Port, R. and Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America*, 66(3):654–662.

Poulios, A. (2009). Age categories as an argumentative resource in conflict talk: evidence from a Greek television reality show. *International Journal of the Sociology of Language*, 200:189–208.

Prince, E. (1987). Sarah Gorby, Yiddish folksinger: A case study of dialect shift. *International Journal of the Sociology of Language*, 67:83–116.

Prince, E. (1988). Accommodation Theory and dialect shift: A case study from Yiddish. *Language and Communication*, 8(3–4):307–320.

Randolph, M. (1989). *Syllable-based constraints on properties of English sounds*. PhD thesis, Massachusetts Institute of Technology.

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

Repp, B. and Williams, D. (1985). Categorical trends in vowel imitation: preliminary observations from a replication experiment. *Speech Communication*, 4(1):105–120.

Reubold, U., Harrington, J., and Kleber, F. (2010). Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication*, 52(7-8):638–651.

Rischel, J. (1992). Formal linguistics and real speech. *Speech Communication*, 11(4-5):379–392.

Robb, M., Gilbert, H., and Lerman, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia phoniatrica et logopaedica*, 57(3):125–133.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite [Computer program].

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

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

Sancier, M. and Fowler, C. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, 25:421–436.

Sankoff, D. and Sankoff, G. (1973). Sample survey methods and computer-assisted analysis in the study of grammatical variation. In *Canadian Languages in their Social Context*, pages 7–64. Linguistic Research, Edmonton, Alberta.

Sankoff, G. (2004). Adolescents, young adults and the critical period: Two case studies from "seven up". In Fought, C., editor, *Sociolinguistic variation: Critical reflections*, pages 121–39. Oxford University Press, New York.

Sankoff, G. (2005). Cross-sectional and longitudinal studies in sociolinguistics. In Ammon, U., Dittmar, N., Mattheier, K., and Trudgill, P., editors, *Sociolinguistics: An international handbook of the science of language and society*, volume 2, pages 1003–13. de Gruyter, Berlin.

Sankoff, G. (2006). Age: Apparent time and real time. In *Encyclopedia of Language and Linguistics*. Elsevier, Oxford. Article Number: LALI: 01479.

Sankoff, G. (2012). Longitudinal studies. Oxford University Press, Oxford. In press.

Sankoff, G. and Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 83(3):560–588.

Sankoff, G. and Blondeau, H. (2010). Instability of the [r]∼[R] alternation in Montreal French: the conditioning of a sound change in progress. Ms., to appear in *'r-atics 2*.

Sankoff, G., Blondeau, H., and Charity, A. (2001). Individual roles in a real-time change: Montreal (r→R) 1947-1995. In van de Velde, H. and van Hout, R., editors, *'R-atics: Sociolinguistic, phonetic and phonological characteristics of /r/*, number 4 in Etudes & Travaux, pages 141–157. ILVP, Brussels.

Santa Ana, O. (1992). Chicano English evidence for the exponential hypothesis: A variable rule pervades lexical phonology. *Language Variation and Change*, 4(3):275–288.

Santa Ana, O. (1996). Sonority and syllable structure in Chicano English. *Language Variation and Change*, 8(1):63–89.

Schilling-Estes, N. (2003). Investigating stylistic variation. In Chambers, J., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 375–401. Wiley-Blackwell.

Schirmunski, V. (1928/1929). Die schwäbischen Mundarten in Transkaukasien und Südukraine. *Teuthonista*, 5(1):38–60, 157–171.

Schirmunski, V. (1930). Sprachgeschichte und Siedelungsmundarten. *Germanistisch Romanistische Monatsschrift*, XVIII:113–122, 171–188.

Schreier, D. (2005). *Consonant change in English worldwide*. Palgrave MacMillan, New York.

Schuchardt, H. (1885). *Über die Lautgesetze. Gegen die Junggrammatiker*. R. Oppenheim, Berlin.

Scobbie, J. (2006). Flexibility in the face of incompatible English VOT systems. In Goldstein, L. M. and Whalen, D.H. Best, C. T., editors, *Laboratory Phonology 8*, pages 367–392. Mouton de Gruyter, Berlin.

Shalev-Shwartz, S., Keshet, J., and Singer, Y. (2004). Learning to align polyphonic music. In *Proceedings of ISMIR-2004*. paper 411.

Shepard, C., Giles, H., and Le Poire, B. (2001). Communication accommodation theory. In Robinson, W. and Giles, H., editors, *The new handbook of language and social psychology*, pages 33–56. Wiley, New York.

Shockey, L. (1984). All in a flap: Long-term accommodation in phonology. *International Journal of the Sociology of Language*, 46:87–95.

Shockley, K., Sabadini, L., and Fowler, C. A. (2004). Imitation in shadowing words. *Perception and Psychophysics*, 66(3):422–429.

Siegel, J. (2010). *Second dialect acquisition*. Cambridge University Press, Cambridge.

Simpson, A. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2):621–640.

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

Smith, J., Durham, M., and Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change*, 21(1):69–95.

Snijders, T. and Bosker, R. (2011). *Multilevel analysis*. Sage, London, 2nd edition.

Sonderegger, M. and Keshet, J. (2010). Automatic discriminative measurement of voice onset time. In *Proceedings of INTERSPEECH-2010*, pages 2242–2245.

Sonderegger, M. and Keshet, J. (2012). Automatic measurement of voice onset time using discriminative structured prediction. Submitted Ms.

Stanford, J. (2007). *Dialect contact and identity: A case study of exogamous Sui clans*. PhD thesis, Michigan State University.

Stanford, J. (2008). A sociotonetic analysis of Sui dialect contact. *Language Variation and Change*, 20(3):409–450.

Steinlen, A. (2005). *The influence of consonants on native and non-native vowel production: a cross-linguistic study*. Gunter Narr Verlag, Tübingen.

Stevens, K. (2000). *Acoustic phonetics*. MIT Press, Cambridge, MA.

Stevens, K. and House, A. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, 6(2):111.

Stone, C. and Koo, C. (1985). Additive splines in statistics. In *Proceedings of the Statistical Computing Section, American Statistical Association*, pages 45–48, Washington, D.C.

Stouten, V. and van Hamme, H. (2009). Automatic voice onset time estimation from reassignment spectra. *Speech Communication*, 51(12):1194–1205.

Stuart-Smith, J. (2004). Scottish English: phonology. In Kortmann and Schneider (2004), pages 47–67.

Summerfield, A. (1975a). Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. *Speech perception*, 2(4). Department of Psychology, Queen's University of Belfast.

Summerfield, A. (1975b). How a full account of segmental perception depends on prosody. In Cohen, A. and Nooteboom, S., editors, *Structure and Process in Speech Perception*, pages 51–68. Springer, New York.

Suomi, K. (1980). *Voicing in English and Finnish stops: A typological comparison with an interlanguage study of the two languages in contact*. University of Turku, Turku.

Swartz, B. (1992). Gender difference in voice onset time. *Perceptual and motor skills*, 75(3):983–992.

Sweeting, P. and Baken, R. (1982). Voice onset time in a normal-aged population. *Journal of Speech and Hearing Research*, 25(1):129–134.

Tagliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge University Press, Cambridge.

Tagliamonte, S. and Baayen, R. (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. Ms.

Tagliamonte, S. and Molfenter, S. (2007). How'd you get that accent?: Acquiring a second dialect of the same language. *Language in Society*, 36(5):649–675.

Tagliamonte, S. and Roberts, C. (2005). So weird; so cool; so innovative: the use of intensifiers in the television series Friends. *American Speech*, 80(3):280–300.

Tagliamonte, S. and Temple, R. (2005). New perspectives on an ol'variable: (t,d) in British English. *Language Variation and Change*, 17(3):281–302.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In Kleijn, W. and Paliwal, K., editors, *Speech coding and synthesis*, pages 495–518. Elsevier, New York.

Taskar, B., Guestrin, C., and Koller, D. (2003). Max-margin Markov networks. In S.Thrun, Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*.

Tauberer, J. (2010). *Learning [voice]*. PhD thesis, University of Pennsylvania.

Thakerar, J., Giles, H., and Cheshire, J. (1982). Psychological and linguistic parameters of speech accommodation theory. In Fraser, C. and Scherer, K., editors, *Advances in the social psychology of language*, pages 205–255. Cambridge University Press, Cambridge.

Theodore, R., Miller, J., and DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America*, 125(6):3974–3982.

Thibault, P. (1991). La langue en mouvement: simplification, régularisation, restructuration. *Linx (Linguistique - Paris X, Nanterre)*, 25:79–92.

Thibault, P. and Daveluy, M. (1989). Quelques traces du passage du temps dans le parler des Montréalais, 1971–1984. *Language Variation and Change*, 1(1):19–45.

Thibault, P., Vincent, D., and Audet, G. (1990). *Un corpus de français parlé: Montréal 84, historique, méthodes et perspectives de recherche*. Départments de langues et linguistique, Université Laval, Québec.

Thornborrow, J. and Morris, D. (2004). Gossip as strategy: The management of talk about others on reality TV show 'Big Brother'. *Journal of Sociolinguistics*, 8(2):246–271.

Tiffany, W. (1959). Nonrandom sources of variation in vowel quality. *Journal of Speech and Hearing Research*, 2(4):305–317.

Tingsabadh, M. and Abramson, A. (1993). Thai. *Journal of the International Phonetic Association*, 23(1):24–28.

Trudgill, P. (1983). Acts of conflicting identity: The sociolinguistics of British pop-song pronunciation. In *On Dialect. Social and Geographical Perspectives*, pages 141–160. Blackwell, Oxford.

Trudgill, P. (1986). *Dialects in contact*. Blackwell, Oxford.

Trudgill, P. (1988). Norwich revisited: Recent linguistic changes in an English urban dialect. *English World-Wide*, 9:33–49.

Trudgill, P. (2004). *New-dialect formation: The inevitability of colonial Englishes*. Oxford University Press, Oxford.

Trudgill, P. (2008). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, 37(2):241–254.

Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the 21st International Conference on Machine Learning*, pages 104–112.

Vallabha, G. and Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. *Journal of the Acoustical Society of America*, 116(2):1184–1197.

van Dommelen, W., Holm, S., and Koreman, J. (2011). Dialectal feature imitation in Norwegian. In Lee and Zee (2011), pages 599–602.

van Hofwegen, J. and Wolfram, W. (2010). Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*, 14(4):427–455.

VanDam, M. and Port, R. (2005). Voice onset time is shorter in high-frequency words. *Journal of the Acoustical Society of America*, 117(4):2623. Poster at `http://www.vandammark.com/docs2/VanDam05_ASAVancouver.pdf`.

Venables, W. and Ripley, B. (2002). *Modern applied statistics with S*. Springer, New York.

Vincent, D., Laforest, M., and Martel, G. (1995). Le corpus de Montréal 1995: Adaptation de la méthode d'enquête sociolinguistique pour l'analyse conversationnelle. *Dialangue*, 6:29–46.

Volaitis, L. and Miller, J. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, 92(2):723–735.

Wagner, S. (2008). *Language change and stabilization in the transition from adolescence to adulthood*. PhD thesis, University of Pennsylvania.

Wagner, S. and Sankoff, G. (2011). Age grading in the Montréal French inflected future. *Language Variation and Change*, 23(3):275–313.

Wahl, A. (2010). The global metastereotyping of ‘Hollywood dudes’: African reality television parodies of mediatized California style. *Pragmatics and Society*, 1(2):209–233.

Walker, J. (2012). Form, function, and frequency in phonological variation. *Language Variation and Change*. In press.

Watson, K. (2007). Liverpool English. *Journal of the International Phonetic Association*, 37(3):351–360.

Watt, D. and Yurkova, J. (2007). Voice onset time and the Scottish vowel length rule in Aberdeen English. In *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 1521–1524.

Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics*, 7:197–204.

Wells, J. (1973). *Jamaican pronunciation in London*. Blackwell, Oxford.

Wells, J. (1982). *Accents of English*. Cambridge University Press, Cambridge. 3 vols.

Whiteside, S. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association*, 26(1):23–40.

Whiteside, S. and Irving, C. (1997). Speakers’ sex differences in voice onset time: some preliminary findings. *Perceptual and motor skills*, 85(2):459–463.

Wikipedia (2012a). Big Brother 9 (UK) — Wikipedia, The Free Encyclopedia. [Online; accessed 21-April-2012].

Wikipedia (2012b). *Big Brother* (UK) — Wikipedia, The Free Encyclopedia. [Online; accessed 21-April-2012].

Wolfram, W. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Center for Applied Linguistics, Washington, DC.

Wood, S. (2012). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.17–13.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, 38(3):329–336.

Yaeger-Dror, M. (1994). Phonetic evidence for sound change in Quebec French. In Keating, P., editor, *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pages 267–293. Cambridge University Press, Cambridge.

Yao, Y. (2009a). An exemplar-based approach to automatic burst detection in spontaneous speech. In Pak (2009), pages 1653–1668.

Yao, Y. (2009b). Understanding VOT variation in spontaneous speech. In Pak (2009), pages 1122–1137.

Zipf, G. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 40:1–95.

Zue, V. (1976). *Acoustic characteristics of stop consonants: A controlled study*. Indiana University Linguistics Club, Bloomington, IN.

APPENDIX A

ANNOTATION

A.1 VOT annotation

All word-initial voiceless stops in the corpus were examined for the presence of a burst. VOT was manually annotated whenever a burst was present, with the exception of stops which were clearly realized as affricates or flaps. I measured a minority of tokens (730), then trained two two undergraduate research assistants, Maria Nelson and Natalie Rothfels, who measured the majority of tokens (896 and 4868).

Whenever possible, VOT was simply taken to be the time difference between the onset of periodicity in the signal and the onset of the burst. A more involved annotation procedure was used in cases where it was not possible to unambiguously identify one or both of these points, as often occurs in conversational speech.

> *Following voiceless segment*: The following vowel is either deleted or present but devoiced, due to the influence of a following voiceless fricative or /h/, for example in reduced *'cause*. In this case, the right VOT boundary was taken to be the boundary between the burst and the following (unvoiced) segment, at the point where there was the most sudden transition from a burst-like spectrum. If there was no abrupt spectral change at the end of the burst, the token was discarded.
>
> *No following segment*: The burst is followed by a period of silence, for example for reduced *to*. In this case the right VOT boundary was taken to be the end of the burst.
>
> *Preceding fricative*: The burst is preceded by a fricative, with no clear stop closure separating them. In this case the left VOT boundary was placed at the point of most rapid spectral change if one could be determined, otherwise simply halfway between the fricative's left boundary and the burst's right boundary.

Double burst: There are two bursts between the closure and the voicing onset, separated by a short silence; usually the later burst is much longer. In this case the left VOT boundary was taken to be the later burst's onset.

A.2 CSD annotation

For each token, three types of annotation were made: the host word's CELEX wordform ID, the realization of the final t/d, and the phonological context of the final t/d. The realization and phonological context were non-trivial, and are described below.

Annotation was performed in two rounds. In the first round, an initial coding of approximately two-thirds of the 6108 tokens was performed by undergraduate and advanced graduate students as the term project for an advanced graduate phonology course.[1] I later did a second pass of the realization and phonological context annotations for all of these tokens. In the second round, two students from the class, Maria Nelson and Natalie Rothfels (both native speakers of American English) underwent additional training with me, then annotated the remaining third of the tokens. We held weekly meetings to ensure consistency of their transcription criteria with mine.

Realization annotation To decide whether and how a word-final coronal was realized, annotators used both acoustic and auditory cues. Emphasis was placed on determining both whether the coronal was perceptibly "there" or not, and whether there were any acoustic cues to realization. Each token's final coronal was classified as one of the following ten labels, corresponding to different phonetic realizations (e.g., burst, glottal stop), the following phonological context, and certainty in the assigned labels.

NONE: Not realized (no perception of t/d presence or clear acoustic cues.)

NONE_BUT: Not realized, but with a caveat (e.g., uncertainty).

[1] The students were Andrea Beltrama, Tasos Chatzikonstantinou, Alastair Cleve, Erin Franklin, Brett Kirken, Jackson Lee, Maria Nelson, Krista Nicoletto, Talia Penslar, Hannah Provenza, and Natalie Rothfels.

BURST: A burst is present which unambiguously belongs to the word-final t/d.

SHARED_COR: There is no burst unambiguously belonging to the word-final t/d, but the next word begins with a coronal stop or affricate (/t/, /d/, /tʃ/, /dʒ/) realized with a burst, which could in principle be "shared" by the t/d and the initial segment of the next word.

SHARED_IDF: The same, but when the next word begins with an interdental fricative (/t/, /ð/) realized as a coronal stop with a burst.[2]

GSTOP: Realization as a glottal stop.

GLOT: Realization as glottalization (but not a full glottal stop).

CLOS: The word-final t/d is realized as an unreleased closure (which is not a glottal stop, i.e., alveolar or dental).

NOCUE, THERE_BUT: Word-final t/d is heard as present, but no acoustic cue can be found or the token is somehow unusual.

Phonological context annotation The two preceding segments and the two following segments were annotated both for their underlying forms and their surface realizations, using the CELEX DISC character set (one character per phoneme). As for realization annotation, annotators used both auditory and spectral cues in deciding on a surface transcription. Because a phonemic character set was used both for the underlying and surface transcription, a few frequent allophones were transcribed as their closest DISC equivalent (i.e., glottal stop as /t/). Pauses following the word-final t/d were also annotated, defined roughly as silences of 30 msec or more. However, we also paid attention to what periods of silence "sounded like" a pause, given the surrounding speaking rate. The end of a speaker's conversational turn was also counted as a pause.

[2] Following Tagliamonte and Temple (2005: 286): "in British English initial interdental fricatives are frequently pronounced as their corresponding stops..."