Title
The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech

Author
Yao, Yao

Publication Date
2011

Peer reviewed|Thesis/dissertation

The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech

by

Yao Yao

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Linguistics

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Keith Johnson, Professor Susanne Gahl, Co-Chairs
Professor Sharon Inkelas
Professor Nelson Morgan

Spring 2011

The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech

by Yao Yao

Abstract

The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech

by

Yao Yao
Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Keith Johnson, Professor Susanne Gahl, Co-Chairs

This dissertation investigates the effects of phonological neighborhoods on pronunciation variation in conversational speech. Phonological neighbors are defined as words that are different in one and only one phoneme by addition, deletion and substitution. Phonological neighborhood density refers to the number of neighbors a certain word has.

Previous research has shown that phonological neighbors impede auditory perception, but facilitate lexical production. As a result, words from dense neighborhoods are harder to perceive, but easier to produce, than words from sparse neighborhoods. Following these effects, two opposite hypotheses can be formed regarding the effects of neighborhood density on pronunciation variation. The listener-oriented hypothesis predicts that high-density words will be hyperarticulated, in order to compensate for their perceptual difficulty. But the speaker-oriented hypothesis predicts that high-density words are more likely to undergo speech reduction, as other words that are easy to produce (such as high-frequency words). To test these hypotheses, two statistical models are constructed to investigate neighborhood effects on pronunciation variation in CVC monomorphemic content words. All speech data come from the Buckeye Corpus of Conversational Speech (Pitt, Dilley, Johnson, Kiesling, Raymond, Hume, and Fosler-Lussier 2007), which contains audio recordings of 40 speaker's free-form interviews. The dataset for the current research consists of more than 500 target words, represented by over 13,000 tokens. The outcome variables in the two models are word duration and degree of vowel dispersion, respectively. Neighborhood density and average neighbor frequency are the critical predictors in both models. Other factors that might affect pronunciation variation in conversational speech, such as word frequency and predictability, are entered into the models as control factors.

The results of the model analyses show that everything else being equal, high-density words are realized with *shorter* durations and *more centralized* vowels than low-density words. These findings provide strong evidence for the speaker-oriented hypothesis. A peripheral finding in the current research is a tendency for words with high-frequency neighbors to have more dispersed vowels. However, this effect is only significant when neighbor frequency from the Hoosier mental lexicon (Nusbaum, Pisoni, and Davis 1984) is used, but not when

neighbor frequency computed from the CELEX database (Baayen, Piepenbrock, and Rijn 1993) is used.

Overall, major findings of the current research support the speaker-oriented hypothesis over the listener-oriented hypothesis, which suggests that word-level pronunciation variation is more heavily influenced by features of the speaker's own production system than by the consideration of listeners' needs.

Professor Keith Johnson, Professor Susanne Gahl
Dissertation Committee Co-Chairs

Contents

List of Figures

List of Tables

Acknowledgments

It has been a long way. At many points, I thought it was hopeless and I should just give up. I am glad that I didn't. More importantly, I feel extremely fortunate to have known some of the greatest people in the world, who supported me to keep going, even in times of setbacks. I will always feel indebted to my thesis committee: Susanne Gahl, Keith Johnson, Sharon Inkelas and Nelson Morgan, without whom the completion of this thesis would be impossible. Susanne and Keith did much more than what is usually expected from thesis committee chairs. They not only provided professional feedback for my work, but also guided me through the process of growing into an independent researcher. Most importantly, they stood by me when I needed support. I have benefited a great deal from their teaching and advising, for which my gratitude cannot be expressed in words. Sharon and Morgan have also played a critical role in helping shape the final product of this research. I am grateful for their effort of reading all the manuscripts (some of which are not the most delightful readings) and providing extremely helpful comments.

In the past several years, I have learned a great deal from my teachers at Berkeley: Ian Maddieson, Line Mikkelsen, Larry Hyman, Leanne Hinton, Andrew Garrett, Gary Holland, just to name a few. Thank you all for kindly sharing with me your knowledge and experience. I would also like to thank my academic brothers, Charles and Maziar; staff in the linguistics department, especially Belén, without whom I would not have survived graduate school; members of the Berkeley phonology lab, especially Reiko, Kiyoko, Ron, Te-hsin, Shira, Sam, Melinda and Molly; other fellow graduate students in the linguistics department, especially Iksoo, Yuni, Michael, Hannah and Erin; my linguist friends outside Berkeley, especially Jackie, Becky, Zhiguo and Pei-Jung. Thank you all for being awesome colleagues and friends.

My gratitude also goes to my family in Shanghai and my dearest non-linguist friends, especially Xiaojing, Xiaoju, Qingfang, Suowei, Xuechun, Li Rong, Cecilia Chu, Qiuyu, Jan and Weigang. Thank you all for keeping me a sane person and showing me how wonderful life is. I am so lucky to know you.

A special person that I owe thanks to is William Shi-Yuan Wang, who opened the world of linguistics to me seven years ago and triggered a wonderful adventure. Now that one adventure is completed, I am looking forward to the next.

Chapter 1

Introduction

We probably all know someone in life who is both a friend and an enemy. It could be a sibling, with whom you have constantly fought for toys in childhood and later for parents' favor. Or a colleague, whom you consider as a good collaborator, but you couldn't help feeling a little annoyed when knowing that they have applied for the same position you want. These people are not real enemies - in fact, you may find them likable in many occasions, but they often instill an uneasy feeling in you because they bring competition. The same relationship may exist among countries or schools, as you learn not to praise France when you are in England or glorify the Stanford campus when you are at Berkeley.

This is a dissertation about the same friend-and-enemy relationship among words in the lexicon. Just like people and countries, words can enter this complicated relationship, too. One of the places where this relationship has been found is among words that sound similar to each other, such as *cat* and *cap*. Previous research has shown that similar-sounding words, or *phonological neighbors* in the technical term, suppress each other in auditory perception (e.g. Luce and Pisoni 1998) but contribute to each other's activation in speech production (e.g. Vitevitch 2002). Put in another way, similar-sounding words are enemies in perception but friends in production (Dell and Gordon 2003). As a result, words from dense neighborhoods, which have many similar-sounding neighbors, are harder to perceive but easier to produce, compared with words from sparse neighborhoods.

In this dissertation, I will investigate how this friend-and-enemy relationship may affect the pronunciation of individual words. Two trends have been observed in previous research on pronunciation variation. On one hand, words that are hard to perceive tend to be hyperarticulated, presumably for the purpose of promoting speech intelligibility. On the other hand, words that are easy to produce are often reduced in speech, as a result of practiced articulatory routines or higher lexical activation. Following these lines, two hypotheses are formulated in this dissertation. A speaker-oriented hypothesis predicts that words from dense neighborhoods should be reduced in speech because they are easy to produce. Contrarily, a listener-oriented hypothesis predicts that high-density words should be hyperarticulated because they are hard to perceive. Thus, the central topic in this dissertation can be related to broader issues regarding the relative strengths of speaker- and listener-oriented forces in

shaping speech.

1.1 Position in the literature

Several previous attempts have been made in the laboratory to study the effects of phonological neighborhoods on word pronunciation (Kilanski 2009; Munson and Solomon 2004; Munson 2007; Scarborough 2004; Wright 1997, 2004). Most of these studies used word-reading experiments, with the exception of Scarborough (2004), which used a modified map task. The most robust finding from this pile of research is that words from dense neighborhoods tend to have more dispersed vowels, which suggests hyperarticulation in high-density words and therefore provides evidence for the listener-oriented account.

A significantly different approach is adopted in this research. Instead of conducting well-controlled speech experiments, I use speech data from the Buckeye Corpus of Conversational Speech (Pitt et al. 2007), which contains 40 native English speakers' speech during free-form interviews. Neighborhood effects on pronunciation variation are studied in two mixed-effects linear regression models, one on word duration, and the other on the degree of vowel dispersion, while other factors that might affect pronunciation variation, such as word frequency and speech rate, are statistically controlled for.

Compared with previous experiments, the dataset in the current research is much larger, with more than 500 words, represented by more than 13,000 tokens from 40 speakers. Both neighborhood density (i.e. the number of neighbors a word has) and neighbor frequency (i.e. the average frequency of neighbors) are coded as continuous variables, and entered into the models as separate factors.

To ensure the validity of the modeling results, both models are tested with several model evaluation techniques for mixed-effects models. Furthermore, a number of alternative models, which utilize different ways of coding neighborhood metrics, are used to test the robustness of model findings.

To preview the results, the major finding from current research is that everything else being equal, words from dense neighborhoods are produced with *shorter* durations and *more reduced* vowels than words from sparse neighborhoods. These findings provide strong evidence for the speaker-oriented hypothesis.

1.2 Outline of the dissertation

The rest of the dissertation is organized as follows. Chapter 2 reviews the relevant literature on phonological neighborhoods and pronunciation variation. It then motivates two sets of predictions (speaker-oriented and listener-oriented) for neighborhood effects on pronunciation, based on the literature review. Chapter 3 contains a detailed description of the corpus and the compilation of the word/token database. It also describes the statistical modeling technsiques (i.e. mixed-effects models) used in the current research. The following two chapters present the results from the corpus studies: Chapter 4 on the model(s) on word

duration, and Chapter 5 on the model(s) on vowel dispersion. The final chapter, Chapter 6, discusses the theoretical implications of the current findings and compares them with previous experimental results.

Chapter 2

General Background

The central question of this thesis is *how does phonological neighborhood structure influence pronunciation variation.* In this chapter, I will formulate two sets of hypotheses related to this question, based on a review of previous literature on phonological neighborhoods and pronunciation variation.

The path of the literature review is as follows: First, I will examine the experimental records for the effects of neighborhood structure on speech perception and production (Section 2.1). The result of this review shows that phonological neighbors inhibit perception but facilitate production. However, relatively little is known about the effects of neighborhood structure on pronunciation variation, much less how to account for them.

To shed some light on the problem, I then survey the broader literature on pronunciation variation (Section 2.2), for potential links from perception and production efficiency to pronunciation variation. The survey reveals two recurring patterns which are largely confluent. One pattern suggests that the degree (or likelihood) or hyperarticulation is positively correlated with the difficulty of comprehension. The other pattern suggests that the degree (or likelihood) or hyperarticulation is positively correlated with the difficulty of production. In many variation phenomena, both correlations are exhibited.

These two correlational patterns lead to two opposite hypotheses (Section 2.3) for the current research question. The perception-based (i.e. "listener-oriented") hypothesis predicts that high-density words should be hyperarticulated, whereas the production-based (i.e. "speaker-oriented") hypothesis predicts that words from dense neighborhoods should be reduced in speech. Both hypotheses are built on the basis of previous findings, and will be tested by the corpus studies presented in Chapters 4 and 5.

2.1 Phonological neighborhoods

2.1.1 Overview

A phonological neighborhood refers to the set of words that sound similar to a given word. The investigation of phonological neighborhoods can be dated to Landauer and

Streeter (1973), but a theoretical framework for discussing neighborhood effects was first laid out by Luce and colleagues (Luce 1986; Luce and Pisoni 1998). In a classic study, Luce and Pisoni (1998), the authors hypothesized that listeners' performance in auditory word recognition is a function of target word frequency and the overall confusability and usage frequency of other similar-sounding words (i.e. phonological neighbors) in the lexicon. High-frequency words will be easier to recognize but words with a large number of phonological neighbors (especially high-frequency neighbors) will be relatively hard to perceive, due to lexical competition. The hypothesis was borne out by experimental results in Luce and Pisoni (1998), as well as a rigorous body of subsequent studies (e.g. Luce, Goldinger, Auer, and Vitevitch 2000; Sommers and Danielson 1999; Vitevitch and Luce 1999; Vitevitch, Stamer, and Sereno 2008).

Later on, the investigation of neighborhood effects has also been extended to the realm of speech production (e.g. Munson and Solomon 2004; Vitevitch 2002; Vitevitch, Ambrüster, and Chu 2004; Wright 1997). Though not entirely uncontroversial, there is mounting evidence that phonological neighbors *facilitate*, rather than *inhibit*, lexical production. Nonetheless, what remains unclear is how phonological neighbors affect the final product of production, i.e. phonetic realization. Proposals have been made, attributing neighborhood-conditioned phonetic variation to ease/difficulty of perception and/or production. However, none of the existing proposals is completely satisfying.

In the rest of this section, I will review the literature on neighborhood effects on both perception and production, with an emphasis on the English language and studies with normal subject populations. Over the past decade or so, while most of the research in the field is still targeted at monosyllabic English words and processing by normal populations, the scope of this research has been significantly extended to words of longer lengths (Cluff and Luce 1990; Vitevitch et al. 2008), languages other than English (Cantonese (Kirby and Yu 2007); Mandarin (Tsai 2007); Spanish (Baus, Costa, and Carreiras 2008; Vitevitch and Stamer 2006, 2009); Hawaiin and Basque (Arbesman, Strogatz, and Vitevitch 2010)), and processing in various subpopulations including children (Coady and Aslin 2003; Garlock, Walley, and Metsala 2001; Hollich, Jusczyk, and Luce 2002; Storkel 2002, 2004, 2006; Woodley 2010), nonnative and bilingual speakers (Marian and Blumenfeld 2006; Marian, Blumenfeld, and Boukrina 2008; Yoneyama and Munson 2010), and patients with language disorders (e.g. Dirks, Takayanagi, Moshfegh, Noffsinger, and Fausti 2001; Sommers, Kirk, and Pisoni 1997). It has been shown that neighborhood effects may vary across languages, even in direction, due to language-specific features in morphology and phonology (Vitevitch and Stamer 2006, 2009). On the other hand, neighborhood research regarding special populations is often motivated by theoretical questions in language acquisition and treatment of language deficiencies. Therefore, for the purpose of a more concentrated discussion, I will limit the current review to the central cluster of the neighborhood literature, which mainly regards lexical processing in English by normal populations.

The organization of the section is as follows. I will first introduce the working definition of phonological neighborhood and neighborhood metrics in §2.1.2, and then review the empirical work on neighborhood effects on perception (§2.1.3), production efficiency (§2.1.4)

and phonetic variation (§2.1.5).

2.1.2 Definition and metrics

2.1.2.1 Definition

Since the investigation of phonological neighborhoods first started in the field of speech perception, the concept of phonological neighborhood is based on sound similarity. A phonological neighbor is a word that *sounds* similar to a given word[1]. Thus, if we consider the mental lexicon as a network of words, phonological neighborhoods can be formed by connecting similar-sounding phonological neighbors.

What is much harder to define is the exact meaning of sound similarity. Obviously some words (e.g. *cat* and *cap*) sound more similar than others (e.g. *tree* and *elephant*). But how to measure similarity and how similar is enough for being considered as neighbors are both tricky questions.

Luce and Pisoni's original experiment (Experiment 1 of Luce and Pisoni 1998, see also Goldinger, Luce, and Pisoni 1989) employed a probabilistic measure of word confusability, in the context of word identification against noise. Instead of directly measuring the similarity between two forms, Luce and Pisoni evaluated the likelihood of confusing the acoustic signal of one form for another in the presence of noise. The more similar two words sound, the more confusable their acoustic signals are. A later study by Yoneyama (2002) employed an acoustic/auditory neighborhood similarity measure based on overall spectral similarity.

However, the use of the above similarity measures is not widespread in subsequent research, possibly due to their computational complexity. Instead, most existing studies on phonological neighborhoods adopted a much simpler definition of phonological similarity using the so-called *one-phoneme difference rule*. Under this definition, any two words that differ in only one phoneme by addition, deletion or substitution are considered as phonological neighbors (Greenberg and Jenkins 1967; Landauer and Streeter 1973; Luce and Pisoni 1998). This definition clearly makes some strong assumptions. For one thing, it assumes that neighborhood membership is a binary property which is either 0 or 1. Second, it considers all phonemes as equally different from one another and all positions in the word as equally important in distinguishing sound sequences. Following from this definition, *cap* and *kit* are equally good neighbors of *cat*, though most people might agree that *cap* sounds more similar to *cat* than *kit* does.

Despite all the simplification, neighborhood metrics generated under the one-phoneme difference rule have proved to be as effective as some more sophisticated measures of perceptual confusability (e.g. Luce and Pisoni 1998)[2]. Thus the use of the one-phoneme difference rule has become the standard in the field. With this definition, the neighborhood of any

[1]Even in later studies on speech production, this perception-based definition of phonological neighborhoods continued to be used. It is not unlikely that with a different type of neighborhood construction, such as one that is based on articulatory similarity, different results may be obtained.

[2]It should be mentioned, however, that Luce and Pisoni's studies, as well as most existing studies on neighborhood effects, use broad-grained neighborhood metrics, e.g., high density vs. low density, high

given word can be easily computed by comparing pronunciations of the given word and every other word in the lexicon.

Most existing neighborhood studies used neighborhood metrics from the Hoosier mental lexicon (HML; Nusbaum et al. 1984), possibly for the comparability of the results. HML neighborhood metrics are based on pronunciations in the Webster's Pocket Dictionary, which contains about 20,000 lexical entries. Each word is compared with every other word in the dictionary and any words that only differ in one phoneme in pronunciation are considered as neighbors. One may question the legitimacy of this practice because many words in the Webster dictionary may be unfamiliar (or even unknown) to normal speakers, and including these words in the neighborhoods seems ungrounded[3]. A more sophisticated approach is to only consider words that are highly familiar to normal speakers as potential neighbors (e.g. Vitevitch 2002; Vitevitch et al. 2004). But the latter approach is less often taken, and so far, no difference has been reported between these two approaches regarding the effectiveness of the neighborhood measures.

2.1.2.2 Metrics

Three neighborhood metrics are most widely used: neighborhood density (i.e. number of neighbors), neighbor frequency (i.e. average usage frequency of neighbors) and target word frequency. All three metrics are provided in the Hoosier mental lexicon, with both frequency measures based on Kučera and Francis (1967).

In addition, it is also possible to develop, based on the three basic measures, more complicated metrics such as frequency-weighted neighborhood density and relative frequency (i.e. the ratio of target word frequency and neighbor frequency). Very recently, researchers have also started to probe into more sophisticated neighborhood structures, such as clustering coefficient (Chan and Vitevitch 2009, 2010) and spread of the neighborhood (Vitevitch 2007), but the results are preliminary.

2.1.2.3 Correlations with other lexical properties

Neighborhood density has been proposed to correlate with a number of lexical properties, including word length, word frequency and phonotactic probability. Among the three, correlations with word length and phonotactic probability have both been confirmed but the conjectured correlation with word frequency is largely disproven. In the following, I will explain each proposed correlation in more detail. The purpose of the discussion is to show which factors need to be treated with special care in a study of neighborhood density, in order to avoid potential confounds.

It is now clear that in the English language, shorter words tend to reside in denser neighborhoods (Bard and Shillcock 1993; Charles-Luce and Luce 1990; Pisoni, Nusbaum, Luce, and Slowiaczek 1985). According to the HML dictionary, a three-phoneme word has

neighbor frequency vs. low neighbor frequency, which are less sensitive to the accuracy of the measures.

[3]If anything, the vocabulary size of a normal English speaker is probably less than 20,000 lexical entries.

18.2 neighbors on average, but an average four-phoneme word only has 6.8 neighbors and a five-phoneme word has even fewer (1.78) neighbors. However, it is important to note that neighborhood density for extremely short words may be inflated. The working definition of neighbors, i.e. the one-phoneme difference rule, only specifies the *absolute* number of phoneme differences (i.e. 1) between neighbors. Thus, the *proportion* of overlapped phonemes is much lower in shorter words than in longer words. Extremely short one- or two-phoneme words in general have as low as 0% or 50% of phonemes in common with their neighbors, but a three- or four-phoneme word shares at least 67.7% of the phonemes with its neighbors. As a result, words of shorter lengths tend to have more neighbors that are actually *not* confusable, which may give rise to overestimated neighborhood density. This provides a motivation for using longer words over shorter words in neighborhood studies. But there are also some motivations for using shorter words, such as simpler syllabic structure and stress patterns, as well as the wider range of neighborhood density in shorter words. In view of these countering arguments, words of three or four phonemes become the most often used targets in the neighborhood literature.

The correlation between neighborhood density and phonotactic probability is also strong (Vitevitch, Luce, Pisoni, and Auer 1999; Vitevitch et al. 2004). Phonotactic probability (or phonological pattern frequency) refers to the likelihood of having the exact phoneme sequence given the distribution of phonemes in the language. In practice, phonotactic probability is usually measured in terms of position-specific phoneme probability (e.g. the probability of having /k/ in initial position) and biphone probability (e.g. the probability of having /kæ/ in first and second positions) (Jusczyk, Luce, and Charles-Luce 1994). Since the phonological patterns in words from dense neighborhoods are shared by many other words in the lexicon, there is a default correlation between neighborhood density and phonotactic probability. In light of this correlation, it is important to check phonotactic probability as a potential confounding factor in any neighborhood study.

The correlation between neighborhood density and word frequency was first suggested in Landauer and Streeter (1973), but has been largely dismissed since then. Landauer and Streeter compared 50 common and rare words and found that the common words had more neighbors than the rare words. However, this relationship did not hold when larger samples of words were analyzed (Frauenfelder, Baayen, Hellwig, and Schreuder 1993; Pisoni et al. 1985). Frauenfelder et al. (1993) calculated neighborhood metrics for both English and Dutch, based on all words in the CELEX lexical database (Baayen et al. 1993), and only found a weak tendency in both languages for density to correlate with word frequency, lacking statistical significance.

2.1.2.4 Limitation

A major limitation of the current neighborhood definition is the failure to take into account of suprasegmental features. It is still an open question whether differences in stress or pitch contribute to distinguishing sound sequences in the same way as phonemic differences. As a result, the vast majority of the existing literature on phonological neighborhoods in

English have only looked at monosyllabic words, which is a large but still limited part of the English lexicon. An attempt to incorporate suprasegmental features is documented in a preliminary study by Kirby and Yu (2007) on a different language, Cantonese. In Kirby and Yu's study, a tonal difference was treated as a segmental difference when defining phonological neighborhoods.

2.1.3 Neighborhood effects on speech perception

The investigation of neighborhood effects first started in the field of auditory word recognition. Most current models (Cohort, Marslen-Wilson and Welsh 1978; NAM, Luce and Pisoni 1998; PARSYN, Luce et al. 2000; Shortlist, Norris 1994; TRACE, McClelland and Elman 1986) assume that word recognition involves an activation-competition process, in which a set of similar-sounding word candidates are activated and the one that has the highest activation or exceeds the activation threshold the earliest will be selected as the target.

The specific mechanisms for the activation-competition process can vary across models (see Jusczyk and Luce 2002 for an extensive review). A major theoretical distinction concerns the identification of word candidates. The original Cohort theory (Marslen-Wilson and Welsh 1978) assumes the most constrained activation, which only allows forms that are consistent with the revealed part of the acoustic stimulus to be activated at any time point. Thus, as the stimulus unfolds, word candidates that no longer match with the acoustic input will be eliminated and the recognition process ends when there is only one candidate left (i.e. a uniqueness point is reached). A later version of Cohort (Warren and Marslen-Wilson 1987), as well as other models (NAM, PARSYN, Shortlist, TRACE), relaxes the constraint by assuming that structurally similar forms can be activated at any point of the speech stimulus, regardless of whether they match with all the acoustic-phonetic information revealed so far or not. This account of radical activation receives support from evidence from eye-tracking experiments (Allopenna, Magnuson, and Tanenhaus 1998), which showed that rhyming neighbors (e.g. *speaker* for *beaker*) are also activated during early stages of recognition, despite the mismatch in initial segments.

There are further distinctions among the models supporting radical activation (NAM, PARSYN, Shortlist, TRACE, and later Cohort), mostly regarding the mechanisms of lexical competition and levels of representation in the model, but these models make largely identical predictions for neighborhood effects. In the current review and discussion, I will adopt NAM (or Neighborhood Activation Model; Luce and Pisoni 1998) as the theoretical framework, because it is the basis for most neighborhood studies.

As mentioned above, the NAM, along with most other current models, assumes an activation-competition process in word recognition. The key component of the NAM is the word decision unit, which outputs a probability (the Frequency Weighted Neighborhood Probability Rule; FWNPR) that indicates the likelihood of a word candidate being the target word. During the recognition process, every word candidate is associated with a decision unit that constantly updates the corresponding FWNPR. The calculation of FWNPR (see

Equation 2.1) takes into account word frequency ($Freq(S)$), consistency with the acoustic stimulus (i.e. perceptual confusability; $p(S)$), and the sum of activity in the system (i.e. the sum of activation of all other activated words plus the activation of the current word). As a result, high-frequency words with forms that are highly consistent with the acoustic input will be associated with high FWNPR.

$$p(ID) = \frac{Freq(S)p(S)}{Freq(S)p(S) + \sum_{j=1}^{n} Freq(N_j)p(N_j)} \tag{2.1}$$

where $p(ID)$ is the probability of identifying the stimulus as the word S, $p(S)$ is the perceptual probability of the word S based on the stimulus input, $Freq(S)$ is the frequency of S, $p(N_j)$ is the perceptual probability of the jth neighbor, and $Freq(N_j)$ is the frequency of the jth neighbor.

Competition among candidates is indirectly implemented via FWNPR. Since the probabilities of all word decision units in the system add up to 1, increase in the FWNPR of one unit will lead to decrease in other units. In case of successful recognition, as the acoustic stimulus proceeds, the probability of the target word will rise and the probabilities of neighbor words will fall, and the recognition process will be completed when the decision unit of the target word outputs a probability that exceeds some threshold.

The NAM generates a number of important predictions. First, it predicts that high-frequency words are easier to recognize than low-frequency words, because the decision units are biased toward high-frequency items. A high-frequency target needs less time to achieve the threshold FWNPR and is less likely to be out-selected by other candidates. Thus, recognition of high-frequency words is both faster and more accurate. Second, the NAM predicts that words with many neighbors are harder to recognize than those with fewer neighbors. For one thing, if the stimulus activates a large set of word candidates, the system will need to keep track of many decision units, which slows down the process. In addition, activating many candidates increases the level of lexical competition in the system, which will result in higher error rates. The third prediction of the NAM is that words with high-frequency neighbors are harder to recognize than those with low-frequency neighbors, because high-frequency neighbors are stronger competitors, due to the frequency bias in the decision units.

The above predictions are shared by the successor of NAM, i.e. PARSYN, and other models of word recognition such as TRACE and Shortlist (though in the latter two models, frequency-bias is encoded in the activation levels of word units and lexical competition is implemented via lateral inhibition), and have been confirmed in perceptual experiments. Evidence for these predictions come from five sources: perceptual identification, lexical decision, auditory naming, same-different tasks and form-based priming. In the following, I will review the results from each experimental paradigm separately.

2.1.3.1 Inhibition from neighbors in perception: Evidence from perceptual identification

In a perceptual identification experiment, subjects are presented with auditory word tokens and are asked to identify the words. Subjects' performance can be evaluated by accuracy of identification. In order to elicit imperfect performance (so that error rates can be analyzed reliably), the stimuli are usually degraded. As discussed above, the NAM predicts that identification accuracy should increase with word frequency, but decrease with neighborhood density and neighbor frequency.

Luce and Pisoni (1998) conducted a perceptual identification experiment (Experiment 1), with about 900 monosyllabic words recorded at different signal-to-noise (SN) ratios (+15 dB, +5 dB and -5 dB). A total number of 90 subjects participated in the experiment. The results showed that identification accuracy was indeed predicted by the recognition probabilities (FWNPR) generated by the NAM. The correlation was higher in high SN conditions ($r>0.4$ at both +15dB and +5dB) than in low SN condition ($r\approx0.2$ at -5dB). Further analysis by block showed that words with high frequency, high signal consistency and low neighbor activity were identified most accurately (mean = 64.03%), whereas words with low frequency, low signal consistency and high neighbor activity were the least accurate (mean = 37.76%).

Subsequent research (Bradlow and Pisoni 1999; Sommers et al. 1997; Sommers and Danielson 1999) confirmed both facilitation from frequency and inhibition from neighborhood activation in perceptual identification of monosyllabic English words. Furthermore, the effects have also been obtained for longer words. Vitevitch et al. (2008) replicated Luce and Pisoni's experiment with 56 bisyllabic word stimuli. All test words were monomorphemic words with strong-weak stress patterns (e.g. *movie*), which according to Cutler and Norris (1988) and Sereno and Jongman (1990), are most common among bisyllabic words in English. Half of the words were from dense neighborhoods (mean = 11.71 neighbors) and the other half were from sparse neighborhoods (mean = 4.43 neighbors). Segmental context and word length (in phonemes) were balanced between high- and low-density words. Similar to monosyllabic words, Vitevitch et al. found that bisyllabic words from dense neighborhoods were identified less accurately (mean = 77.1%) than those from sparse neighborhoods (mean = 80.3%).

2.1.3.2 Inhibition from neighbors in perception: Evidence from auditory lexical decision

In the auditory lexical decision paradigm, subjects are presented with auditory stimuli and are asked to decide, as quickly and accurately as possible, whether the stimuli are real words or not. Subjects' performance is measured by both speed and accuracy of their response.

For a word stimulus, processing in lexical decision is similar to that in perceptual identification: a number of similar-sounding candidates are activated by the stimulus, among

which the one that first exceeds the activation level threshold will be selected and a word response will be given. Thus, like in perceptual identification, the NAM predicts that classification responses to words from dense neighborhoods will be slower than those from sparse neighborhoods. However, neighborhood density may not affect response accuracy, because a misidentified target (which will generate an error in perceptual identification task) will also give a correct word response. For a nonword stimulus, presumably none of the word candidates should have an activation level higher than the threshold, but the NAM predicts that response latency for a nonword stimulus depends on the overall activity in the system. If the overall activation is lower than a low threshold, a nonword response will be quickly given, otherwise, a decision will be enforced at the self-imposed deadline (Coltheart, Davelaar, Jonasson, and Besner 1976). Since a dense neighborhood will increase the overall activity, responses to nonword stimuli with many neighbors will be both slower and less accurate.

The above predictions were confirmed in Experiment 2 of Luce and Pisoni (1998) and later studies by Vitevitch and colleagues (Vitevitch and Luce 1999; Vitevitch et al. 2008). Luce and Pisoni conducted an auditory lexical decision experiment with the same set of word stimuli as in the perceptual identification task (Experiment 1, see the discussion above) and 300 nonword stimuli with comparable lengths and matching phonemic contents. The word stimuli were divided by frequency (high and low), neighborhood density (high and low) and neighbor frequency (high and low), while the nonword stimuli by neighborhood density (high and low) and neighbor frequency (high and low). Thirty subjects participated in the experiment, with each subject listening to an equal number of word and nonword stimuli. Classification responses were given by pressing buttons.

For word stimuli, the effects of frequency and neighbor frequency were clear: high-frequency words (mean latency = 390ms; mean accuracy = 93.43%) were responded to faster and more accurately than low-frequency words (mean latency = 445ms; mean accuracy = 86.03%); words with high-frequency neighbors (mean latency = 426.25ms; mean accuracy = 89.04%) were responded to more slowly and less accurately than words with low-frequency neighbors (mean latency = 408.75ms; mean accuracy = 90.42%). Nevertheless, the effects of neighborhood density were more complicated. High-density words (mean latency = 424.25ms; mean accuracy = 91.4%) were classified more slowly but *more* accurately than low-density words (mean latency = 410.75ms; mean accuracy = 88.04%). Further analysis showed that the density effect on latency only existed in high-frequency words and the density effect on accuracy only existed in low-frequency words. In view of this, Luce and Pisoni argued that low-frequency words are in general hard to recognize, therefore the decisions are made after self-imposed deadlines expire, in which case higher overall activity in the system (as in the case of a dense neighborhood) would bias toward a word response, resulting in higher response accuracy.

For nonword stimuli, both neighborhood density and neighbor frequency affected nonword classification in the expected directions. Nonwords occurring in high-density neighborhoods (mean latency = 451ms; mean accuracy = 86.55%) were responded to more slowly and less accurately than those in low-density neighborhoods (mean latency = 411.5ms; mean accuracy = 90.03%), and the same contrast was found between nonwords occurring in high-

frequency (mean latency = 437ms; mean accuracy = 86.85%) and low-frequency neighborhoods (mean latency = 425.5ms; mean accuracy = 89.74%).

Vitevitch and Luce (1999, Experiment 3) confirmed the above trends. Moreover, Vitevitch et al. (2008, Experiment 2) extended the findings to bisyllabic words, showing that high-density bisyllabic words were responded to more slowly and less accurately than low-density bisyllabic words.

2.1.3.3 Inhibition from neighbors in perception: Evidence from auditory naming

In an auditory naming task, subjects hear spoken stimuli and are asked to repeat the stimuli as quickly and accurately as possible. Subjects' performance can be assessed by both response latency and accuracy. Compared with perceptual identification and auditory lexical decision, auditory naming has two distinctive features. First, a naming response does not require lexical recognition *per se*. It is possible for subjects to optimize performance by concentrating on repeating the acoustic-phonetic patterns in the input without lexical recognition. Second, auditory naming involves an extra module of production - articulation. An articulatory plan must be compiled and executed at the end of each trial. Thus, factors that influence lower-level articulation may also be manifested in naming performance.

Experiment 3 of Luce and Pisoni (1998) showed that neighborhood density had an inhibitory effect on word naming. Luce and Pisoni used 400 CVC words, evenly divided into 8 (high and low frequency × high and low density × high and low neighbor frequency) conditions. Results showed that high-density words (mean latency = 822ms; mean accuracy = 97.8%) were named more slowly and slightly less accurately than low-density words (mean latency = 720ms; mean accuracy = 98.08%). However, no effects of word frequency and neighbor frequency were found, except for a small advantage (0.89%) in accuracy for high-frequency words. According to Luce and Pisoni, the presence of the density effect suggests that processing in auditory naming is sensitive to lexical factors, but the overall absence of frequency effects indicates that the decision units are not biased by frequency. Luce and Pisoni further attributed the absence of the frequency bias to a task-specific feature, since the main function of the frequency bias is to optimize lexical recognition, but recognition is not required in auditory naming.

Similar results of inhibition from neighborhoods have been obtained in a speech shadowing experiment in Vitevitch and Luce (1998), which found that high-density words had longer latencies than low-density words.

One may argue that these density effects may be due to phonotactic probability, which is correlated with density but more interpretable in terms of lower-level articulation. However, this possibility is not likely to be true. Despite the high correlation, neighborhood density and phonotactic probability have been shown to have opposite effects on auditory recognition (e.g. Luce and Large 2001; Vitevitch and Luce 1998, 1999; see the discussion below under "Evidence from same-different tasks"). While high neighborhood density *inhibits* recognition, high phonotactic probability actually *facilitates* recognition. Thus, if the observed density

effect in Experiment 3 of Luce and Pisoni (1998) was indeed owing to phonotactic probability, the direction of the effect should be facilitative rather than inhibitory.

2.1.3.4 Inhibition from neighbors in perception: Evidence from same-different tasks

In a same-different experiment, subjects are presented with two successive spoken stimuli, and are asked to indicate whether the stimuli are the same or different. Similar to auditory naming, a same-different task does not require processing at the lexical level, because the task can be accomplished simply by comparing the acoustic-phonetic patterns of the stimuli. If the stimuli are real words, a lexical effect may be unavoidable as corresponding lexical units may still be activated. The NAM predicts that as long as there is processing at the lexical level, neighborhood density will generate an inhibitory effect.

The prediction was borne out by experimental evidence (Luce and Large 2001; Vitevitch and Luce 1999). Vitevitch and Luce (1999) conducted a speeded same-different experiment with 140 word and 240 nonword CVC stimuli. Both word and nonword stimuli were divided into two subgroups: the high density/phonotactic probability group and the low density/phonotactic probability group. Frequency, uniqueness point and initial segments were balanced between the two subgroups in both word and nonword stimuli. 18 subjects participated in the experiment, and their responses to the "same" stimuli pairs were analyzed. The results showed that for real words, an expected inhibitory effect from neighbors was obtained. Words of high density/phonotactic probability (mean = 972ms) were responded to more slowly than words of low density/phonotactic probability (mean = 926ms). But for nonwords, a reverse pattern was observed, as high-density/phonotactic probability stimuli (mean = 1055ms) were responded to faster than low-density/phonotactic probability stimuli (mean = 1102ms).

Vitevitch and Luce interpreted the results as showing differences between processing at the lexical level and the sublexical level. While lexical recognition is inhibited by phonological neighbors (as shown in the word stimuli), sublexical processing is facilitated by high-frequency phonological patterns (as shown in the nonword stimuli). This is in line with previous evidence from a speech shadowing experiment by the same researchers (Vitevitch and Luce 1998), which showed that density/phonotactic probability had an inhibitory effect on the shadowing of word stimuli, but a facilitative effect for nonword stimuli. Vitevitch and Luce (1999) further demonstrated that the level of representation involved in processing can be manipulated by modifying the task. If the task is biased toward using a lexical processing strategy (e.g. when the word stimuli are presented by block), the results would exhibit an inhibitory effect of density/phonotactic probability; if the task is biased toward using sublexical processing (when word and nonword stimuli are presented in a mixed order), the results would exhibit a facilitative effect of density/phonotactic probability.

More direct support for two levels of processing comes Luce and Large (2001), who disentangled (frequency weighted) neighborhood density and phonotactic probability with the same paradigm. Luce and Large orthogonally combined two levels (high and low) of each

variable in the stimuli set[4]. Forty-five stimuli were selected for each of the eight (density × phonotactic probability × lexicality) conditions. The analysis of 45 subjects' correct "same" responses showed that for word stimuli, there were simultaneously an inhibitory effect of neighborhood density and a facilitative effect of phonotactic probability. High-density words (mean = 708ms) were responded to slower than low-density words (mean = 688ms), but high-phonotactic probability words (mean = 689ms) were responded to faster than low-phonotactic probability words (mean = 707ms). For nonword stimuli, however, no effect of neighborhood density or phonotactic probability (either main effect or interaction) was found. Luce and Large attributed this failure to inaccurate estimation of lexical competition (by neighborhood density) for nonwords in the low-density and high-phonotactic probability condition.

Overall there is strong evidence that responses to word stimuli in a same-different task are inhibited by neighborhood density, as predicted by the NAM.

2.1.3.5 Inhibition from neighbors in perception: Evidence from form-based auditory priming

Experimental paradigms that are discussed above can also be combined with auditory priming. When priming is used, subjects are exposed to an auditory token of a prime stimulus first, before the perceptual task. Thus, performance on the subsequent perceptual task may exhibit influences from the prime stimulus. The NAM predicts that if the prime stimulus activates a phonological neighbor of the target (as in form-based priming), target recognition will be inhibited, as compared to when the prime activates an unrelated word.

Evidence regarding this prediction is not entirely consistent. Previous research has obtained both facilitation (Radeau, Morais, and Dewier 1989; Slowiaczek and Pisoni 1986; Slowiaczek, Nusbaum, and Pisoni 1987) and inhibition (Goldinger et al. 1989; Goldinger, Luce, Pisoni, and Marcario 1992) with form-based priming. Slowiaczek et al. (1987) observed that perceptual identification was more successful when targets were preceded by primes that shared the initial phonemes (e.g. *bat-bin*) than when preceded by unrelated primes, but no effect was observed in auditory lexical decision (Radeau et al. 1989; Slowiaczek and Pisoni 1986). On the other hand, Goldinger et al. (1989) used primes that were phonetically similar but phonemically non-overlapping with the targets (e.g. *pat-bin* and *veer-bull*), and obtained an inhibitory effect on subsequent target identification, though the effect was only robust with short interstimulus intervals and low-frequency primes.

An obvious difference between these two sets of findings is that they relied on different types of form-based priming. As termed by Goldinger et al. (1992), one is "phonological priming", in which the prime and target share phonemes (e.g. *bat-bin*), and the other is "phonetic priming", in which the prime and target are phonetically similar but do *not* share phonemes (e.g. *pat-bin*). Goldinger et al. (1992) replicated both facilitation of phonological priming and inhibition of phonetic priming as found in previous studies. More importantly,

[4]Because of the high correlation, Luce and Large's stimuli had smaller ranges of variation in absolute density and phonotactic probability, compared to Vitevitch and Luce's stimuli.

Goldinger et al. showed that when the proportion of related primes was reduced from 50% to 10% (i.e. unrelated primes increased from 50% to 90%), the effect of phonemic priming was significantly reduced (in word identification) and even reversed (in lexical decision), but the inhibitory effect of phonetic priming remained unchanged. This suggested that the observed facilitation of phonemic priming may be artifactual of a guessing strategy employed by the subjects, biasing the identification responses to phonemically similar forms (especially in the initial segments) based on a speculated relationship between prime and target. The results with phonetic priming, on the other hand, appeared to reflect true inhibitory properties of similar-sounding neighbors in spoken word recognition.

Thus we can say that as predicted by the NAM, priming with form-related items inhibits lexical recognition, though the effect is sometimes masked by task-specific strategies.

2.1.3.6 Summary of inhibitory effects of phonological neighbors in perception

To recapitulate, previous research has shown that similar-sounding neighbors inhibit lexical recognition, even when listeners are not exposed to the neighbors in the course of the experiment. Evidence comes from a variety of experimental paradigms. As predicted by the NAM and other compatible models of auditory word recognition (TRACE, Shortlist, PARSYN, etc), the difficulty of lexical recognition increases with both neighborhood density and neighbor frequency, and the effects are attested for both monosyllabic and bisyllabic words.

It is important to note that so far, I have only reviewed the evidence for word recognition in isolation. Neighborhood effects in context have also been examined, though to a lesser degree. Sommers and Danielson (1999) investigated performance on perceptual identification under three context conditions: single words, low-probability sentence contexts and high-probability sentence contexts. Sommers and Danielson found an overall inhibitory effect of phonological neighborhoods, but the size of the effect was reduced in high-probability contexts. Along the same line, Sommers et al. (1997) showed that the inhibitory effects were reduced when subjects were asked to choose from a closed set of candidates in identification tasks.

A more recent study by Goldwater and colleagues (Goldwater, Jurafsky, and Manning 2010) provided some insights from a different perspective. Goldwater et al. analyzed recognition errors made by state-of-the-art speech recognizers, and found that words with one or two strong competitors which were both contextually plausible and acoustically confusable were particularly hard to recognize. This finding also suggests that there might be an interactive effect of contextual predictability and neighborhood density on the efficiency of speech recognition.

Another point that is worth mentioning (again) is the relationship between neighborhood density and phonotactic probability. Despite the high correlation (Vitevitch et al. 1999), previous research has successfully unconfounded the effects of these two variables for word recognition. While dense neighborhoods inhibit lexical recognition, high-frequency phonological patterns facilitate sublexical processing (Luce and Large 2001; Vitevitch and

Luce 1998, 1999). When the two variables are contrasted, both effects can be observed simultaneously (though they may not operate in an additive manner, see Luce and Large 2001). When the two variables are merged, direction of the surface effects depends on the mode of processing (i.e. lexical or sublexical) preferred by the task (Vitevitch and Luce 1999). Given this evidence, it is not possible to attribute the observed inhibitory effects of neighborhood density to phonotactic probability.

2.1.4 Neighborhood effects on production efficiency

Despite the inhibitory effects of phonological neighborhoods on word recognition, there is increasing evidence that phonological neighbors play an opposite role in speech production. Words from dense neighborhoods are produced with greater efficiency (i.e. shorter latency and higher accuracy) than those from sparse neighborhoods. Evidence comes from naturalistic and elicited speech errors and picture naming studies.

2.1.4.1 Facilitation from neighbors in production: Evidence from malapropism

A malapropism is a speech error that involves substituting an error word for a similar sounding target (e.g. saying *octane* as *octave*). Vitevitch (1997) examined 138 malapropisms from a corpus collection of naturalistic speech errors (Fay and Cutler 1977). A comparison of neighborhood characteristics with 10 random samples (of words of comparable lengths and syntactic categories) from an English lexicon revealed that malapropism targets had significantly lower frequency and lower neighborhood density and neighbor frequency. That is to say, low-frequency words and words from sparse neighborhoods are more susceptible to word substitution errors than high-frequency words and words from dense neighborhoods.

2.1.4.2 Facilitation from neighbors in production: Evidence from TOT elicitation tasks

A TOT (tip-of-the-tongue) state occurs when the retrieval of a lexical item fails but higher-level syntactic and semantic information is accessible (Brown 1991). In laboratory settings, a TOT state can be elicited using the TOT elicitation technique (Brown and McNeill 1966). In this methodology, subjects are presented with a question with a word definition (e.g. *What do you call a navigational instrument used in measuring angular distances, especially the altitude of sun, moon and stars at sea?*) and are asked to respond with the target word (e.g. *sextant*). If the subject fails to give the correct answer, they can either indicate that they know the word but cannot retrieve it (i.e. TOT state) or that they do not know the word (i.e. unknown target). The critical measure in a TOT elicitation task is the rate of TOT responses.

The TOT elicitation method provides a particularly useful tool for getting insights into the process of phonological form retrieval. In the neighborhood literature, it has been found that the existence of phonological neighbors, whether or not explicitly invoked in context,

induces lower TOT rates in production. A study by James and Burke (2000) showed that fewer TOT states were elicited when subjects were asked to read aloud form-related primes before each trial. A seemingly contradictory finding was reported in Jones and Langford (1987), who found that more TOT states were associated with form-related auditory priming than with unrelated primes. However, as later studies (Meyer and Bock 1992; Perfect and Hanley 1992) have pointed out, Jones and Langford' stimuli differed across conditions in their susceptibility to TOT states.

More direct evidence for neighborhood facilitation comes from Harley and Bown (1998) and Vitevitch and Sommers (2003), both of which showed that words from dense neighborhoods were less likely to generate TOT responses than those from sparse neighborhoods, even when the neighbors were not explicitly presented. Vitevitch and Sommers's stimuli included 120 CVC target words that were evenly divided into 8 (high and low frequency $\times$ high and low density $\times$ high and low neighbor frequency) conditions. The results revealed main effects of both frequency and neighborhood density: fewer TOT states were elicited with high-frequency words (mean = 1.3%) than with low-frequency words (mean = 2.9%), and with high-density words (mean = 1.2%) than with low-density words (mean = 3%). But no effect of neighbor frequency was observed.

Overall, evidence from TOT elicitation experiments suggests that phonological neighbors help prevent TOT states.

2.1.4.3 Facilitation from neighbors in production: Evidence from tongue twister tasks

A tongue twister is composed by juxtaposing words that contain highly similar sounds (e.g. *twin-screw steel cruiser*), which are therefore hard to articulate. In a tongue twister experiment (Shattuck-Hufnagel 1992), subjects are asked to repeat constructed tongue twisters as quickly as possible for a number of times. The critical measure in this experiment is the rate of erroneous pronunciation of critical segments.

Vitevitch (2002, Experiment 2) showed that tongue twisters made with high-density words were easier to pronounce than those made with low-density words. Each of the 20 tongue twisters used in Vitevitch's experiment consisted of four CVC words with highly confusable initial phonemes (e.g. *peach balm bull pig*). Half of the tongue twisters contained high-density words and the other half contained low-density words. Frequency and neighbor frequency were balanced across conditions. To prevent any articulatory differences, high- and low-density tongue twisters were also matched in initial phonemes. During the experiment, subjects were asked to repeat each tongue twister six times as quickly as possible. Results from 28 participants showed that high-density tongue twisters induced fewer errors (mean = 7%) than low-density stimuli (mean = 12%), suggesting that high-density words are less likely to be affected by competing articulatory plans.

2.1.4.4 Facilitation from neighbors in production: Evidence from SLIP tasks

The SLIP task (Spoonerisms of Laboratory Induced Predisposition, Baars 1992; Baars and Motley 1974; Motley and Baars 1976) is a laboratory method for eliciting spoonerisms, which is a type of speech error (or deliberate play on words) involving swapping segments between words, e.g. *flight the liar* for *light the fire*. The idea of this methodology is to prime subjects with competing speech plans, which may induce spoonerisms in target word production. The experiment proceeds as follows: the subject is visually presented with a list of word pairs and is asked to repeat to themselves after each pair; occasionally the subject is cued to read a word pair (i.e. target pair) out loud. Word pairs that precede the target pairs (i.e. prime pairs) usually serve as distractors. For example, after practicing *bull-pit* and *book-piss*, both of which would prime a /b/-/p/ speech plan, the subject may be asked to read out *push-big*, which requires a /p/-/b/ plan. Subjects' performance is evaluated by error rates in target pair production.

Vitevitch (2002, Experiment 1) showed that words from dense neighborhoods are less likely to elicit spoonerisms than those from sparse neighborhoods. A hundred and twelve target word pairs were used inVitevitch's study, with each pair consisting of two CVC words that were matched in frequency and neighborhood characteristics. The target pairs were evenly divided among eight (high and low frequency × high and low density × high and low neighbor frequency) conditions. During the experiment, each target pair was preceded by three distractor pairs. Results from 78 participants revealed main effects of both frequency and neighborhood density, though no effect was found for neighbor frequency. More speech errors were elicited for low-frequency pairs (mean = 31.9%) than for high-frequency pairs (mean = 16.4%), and for low-density pairs (mean = 31.6%) than for high-density pairs (mean = 16.8%). This evidence suggested that the production of high-frequency words and high-density words is less likely to be interfered by competing speech plans than that of low-frequency words and low-density words.

Similar facilitative effects were obtained in Stemberger (2004), with a reanalysis of results from a previous SLIP experiment (Stemberger and Treiman 1986; Stemberger 1991). More specifically, Stemberger argued that the facilitative effects were due to the gang of *friends* among phonological neighbors. The gang of *friends* is defined by Stemberger as the set of neighbors "that share a particular characteristic with the target word and reinforce it" (p416); by contrast, there is also a gang of *enemies* that does not share the particular characteristic with the target word. According to Stemberger, the larger the gang of friends is, the more the critical characteristic is reinforced and the lower the error rate is. The effect of the gang of enemies, on the other hand, depends on both the size of the gang and the coherence of the gang (i.e. the coherence in proposed competitors). A high-coherence gang of enemies will produce an inhibitory effect by activating some strong competitor, but a low-coherence gang of enemies will have little or no effect on processing.

Stemberger's proposal provides a way for explaining the overall positive effects of neighborhood density in preventing spoonerisms and other types of speech errors. Because density

is probably positively correlated with the size of the gang of friends[5], and coherence of the gang of enemies is usually low, one would expect that an increase in neighborhood density will lead to more facilitation than inhibition and therefore the overall effect is facilitative.

2.1.4.5 Facilitation from neighbors in production: Evidence form picture naming

In a picture naming task, subjects are presented with a picture or drawing and are asked to name the illustrated word (i.e. picture name) as quickly and accurately as possible. There is usually a familiarization procedure before the task, in which subjects learn about and remember the picture names. Performance on picture naming can be evaluated by naming latency and accuracy.

It should be noted that a picture naming response involves several subtasks, including picture recognition, lexical access and articulation, therefore naming performance may be subject to factors that influence any of these subtasks.

A number of studies have shown that naming response is facilitated when preceded by form-related auditory primes, compared with unrelated primes (Costa and Sebastian-Galles 1998; Jescheniak and Schriefers 2001; Meyer 1996; Schriefers, Meyer, and Levelt 1990). More direct evidence for facilitation from phonological neighborhoods comes from naming studies without priming. Vitevitch (2002) presented a series of experiments (Experiment 3, 4 and 5) that examined the effects of neighborhood density on picture naming efficiency. Both Experiment 3 and 4 are standard naming experiments. Experiment 3 used 48 drawings, half of which illustrated high-density monosyllabic words (mean =19.4 neighbors) and the other half illustrated low-density monosyllabic words (mean = 6.8 neighbors). Frequency, familiarity, neighbor frequency and the distribution of initial phonemes were all balanced across conditions. Results from 34 participants revealed a facilitative effect of neighborhood density: Drawings with names from dense neighborhoods (mean = 716ms) were responded to faster than those from sparse neighborhoods (mean = 739ms). No effect was observed in error rates (with correct responses at about 94% of the time in both conditions), suggesting that subjects did not sacrifice accuracy for speed.

Experiment 5 used a modified picture naming task with nonverbal responses to rule out any articulatory effect. Subjects were asked to press a button as soon as they retrieve the target name, and then say the name aloud (for assessing accuracy). Latency was measured from the response box. The stimuli were the same as in Experiment 4. Responses from 25 subjects showed that latencies were in general shorter than in Experiment 4, but the same facilitative effect of neighborhood density was obtained, with a similar magnitude. Drawings in the high-density condition were responded to about 21ms faster than those in the low-density condition, without sacrificing naming accuracy. This finding is significant, because

[5]Although De Cara and Goswami (2002) has shown that monosyllabic words (in English) of high density tend to have a higher proportion of rhyme neighbors than onset neighbors, compared with word of low density, it is still very likely that overall neighborhood density correlates with the numbers of neighbors at each position.

it suggests that the locus of the density effect is earlier than articulation.

2.1.4.6 Disentangling phonotactic probability and neighborhood density

As mentioned before, there is a default high correlation between neighborhood density and phonotactic probability, because the phonological sequences in high-density words are shared by many other words (i.e. phonological neighbors) in the lexicon. In addition, there is independent evidence showing that high phonotactic probability facilitates sublexical processing in both speech perception (Luce and Large 2001; Vitevitch and Luce 1998, 1999; see the discussion before) and production (Munson 2001; Vitevitch et al. 2004). Therefore, it is important to show that the facilitative effects of neighborhood density are *not* confounded by phonotactic probability. In the past literature, several picture naming experiments have been carried out by Vitevitch and colleagues (Vitevitch 2002; Vitevitch et al. 2004) specifically for this purpose.

Experiment 4 of Vitevitch (2002) unconfounded neighborhood density and phonotactic probability by controlling for phonotactic probability in the stimuli. The author ensured that there was no significant difference in either phoneme probability or biphone probability between high- and low-density target names, while maintaining the balance in frequency, familiarity and neighbor frequency. Furthermore, the distribution of initial phonemes was matched exactly across conditions, so that each initial phoneme was present in the same number of high- and low-density words. Thus, not only was phonotactic probability balanced, there should also be no difference in the sensitivity of the voice key for detecting oral naming responses. Altogether 48 drawings were used in this experiment, half of which in the high-density condition (mean = 21.38 neighbors) and the other half in the low-density condition (mean = 11.72 neighbors). Responses from 25 subjects showed that drawings in the high-density condition were named about 25ms faster than those in the low-density condition, without sacrificing accuracy. The results suggested that neighborhood density has a facilitative effect on picture naming that is beyond the effect of phonotactic probability.

Nevertheless, a later naming experiment by the same researcher (Vitevitch et al. 2004, Experiment 3), which varied neighborhood density and phonotactic probability orthogonally, seemed to have obtained different results. Facilitation of phonotactic probability was confirmed, but no effect of neighborhood density was found. One notable difference is that in the 2004 experiment, the absolute range of variation in neighborhood density is much smaller compared to the 2002 experiment. As shown in Table 2.1, stimuli used in the 2002 experiment were comparable in phonotactic probability to those in the high-probability condition of the 2004 experiment, but exhibited greater differences between high- and low-density conditions than the corresponding condition of the 2004 experiment. More importantly, stimuli in the low-density condition of the 2002 experiment covered a lower density range (mean = 11.72) than those in the low-density and high-probability condition of the 2004 experiment (mean = 16.2). For comparison, the low-density condition in Experiment 3 was even lower, with mean density = 6.8 neighbors. Thus it is possible that the failure to observe any density effect in the 2004 experiment was at least partly due to an overall elevated level of

Study	Phonotactic probability		Neighborhood density	
			High	*Low*
Experiment 4 of Vitevitch (2002)	p_{sing}=0.15; p_{bi}=0.006		21.38	11.72
Experiment 3 of Vitevitch et al. (2004)	*High*	p_{sing}=0.16; p_{bi}=0.008	23.9	16.2
	Low	p_{sing}=0.11; p_{bi}=0.004	23.3	15.3

Table 2.1: Mean phonotactic probability and neighborhood density in each stimulus condition of Experiment 4 of Vitevitch (2002) and Experiment 3 of Vitevitch et al. (2004). p_{sing} is the average position-specific single-phone probability; p_{bi} is the average position-specific biphone probability.

neighborhood density in the stimuli.

Given the evidence from Experiment 4 of Vitevitch (2002), the observed facilitative effects of neighborhood density on speech production are at least not fully confounded by phonotactic probability. This suggests that the locus of the effects is not purely at the sublexical level, which is also corroborated by the fact that the effects are prominent before articulation (as shown by Experiment 5 of Vitevitch (2002)). If we also take into account previous evidence regarding TOT states, it is strongly implied that the locus of the observed facilitative effects of neighborhood density is located in phonological form retrieval.

2.1.4.7 Understanding the facilitative effects of neighborhood density on lexical production

Overall, previous research has shown that phonological neighbors, whether explicitly invoked in context or not, facilitate the production of the target form. Words from dense neighborhoods are produced with shorter latencies and higher accuracy compared with words from sparse neighborhoods. Furthermore, the facilitative effects have been shown to exist in pre-articulation stages and persist after controlling for phonotactic probability. Considering evidence from TOT states, the most likely locus of the neighborhood effects is in phonological form retrieval.

An immediate question is how to interpret such effects in a model of lexical production. As mentioned before, it is widely accepted that lexical production consists of several stages, which correspond to processing at different levels (concepts, lemmas, lexemes and sublexical phonemes). A crucial difference among existing models is the direction of information flow. In strictly feedforward models (Levelt, Roelofs, and Meyer 1999), information can only be passed from an earlier stage to a later stage, but in interactive models (Dell 1986), this constraint is relaxed and both feedforward and feedback are allowed. Since current models of

speech production generally assume no connections among items at the same level, the effects of phonological neighbors have to be interpreted in terms of phonological-lexical interaction. By way of an example, when the phonological form /kæt/ is activated, the activation will be spread to component phonemes (/k/, /æ/, /t/) at the phoneme level, which in turn will be passed back to the lexeme level, activating similar-sounding forms that also contain these phonemes (e.g. /mæt/, /kæp/, /kʌt/, etc). Thus, the presence of neighborhood effects in production provides crucial evidence for interactive theories of speech production over strictly feedforward models.

A second question is how to explain the contrast between speech perception and production regarding neighborhood effects. As reviewed in §2.1.3, speech perception exhibits an opposite, inhibitory effect of neighborhood density. That is to say, words from dense neighborhoods are *harder* to recognize, but *easier* to produce, compared with words from sparse neighborhoods. On a conceptual level, as the title of Dell and Gordon (2003) suggests, neighbors can be both *friends* and *foes*: neighbors are friends because they reinforce the characteristics that are shared with the target; neighbors are foes because they introduce competition by activating characteristics that are not present in the target (Dell and Gordon 2003; Stemberger 2004). Thus, the key question is *why are neighbors friends in production but foes in perception.*

There are two important differences between perception and production. First, perception and production differ in the direction of processing. Auditory perception mostly proceeds from sound to meaning (especially in isolated word perception), but speech production proceeds from meaning to sound. Second, a major goal (and probably the most important goal) of speech perception is to decide which words have been heard, whereas the end result of speech production is articulation. Thus, in speech perception, neighbors are activated early on in the process, based on acoustic similarity, and they compete for final selection throughout the process. By contrast, in speech production, neighbors are activated later in the process, at least after the target form is activated in the stage of lexical access. Furthermore, although the activation of neighbors may introduce competition at the lexeme level, it ultimately reinforces individual segments in the target form and therefore facilitates target production. Thus, on the surface, neighbors will appear to be more of competitors in speech perception, but more of helpers in speech production.

Having established the opposite effects of neighborhood density on perception and production efficiency, I will now turn to the central theme of the current work, i.e. the effects of neighborhoods on phonetic variation. Previous work in this field has ascribed neighborhood effects on phonetic variation to neighborhood effects on perception and production efficiency. In the following section, I will review this part of the literature in detail.

2.1.5 Neighborhood effects on phonetic variation

A number of studies (Baese-Berk and Goldrick 2009; Goldinger and Summers 1989; Kilanski 2009; Munson 2007; Munson and Solomon 2004; Scarborough 2004, in press; Watson and Munson 2007; Wright 1997, 2004) have examined the effects of neighborhood structure

on phonetic variation, especially with regards to VOT, vowel dispersion and coarticulation. Most of these studies suggested some form of phonetic strengthening (e.g. longer VOT and more dispersed vowels) in words from dense neighborhoods, though the results are not entirely coherent. The interpretation of these results in the previous literature heavily relied on neighborhood effects on perception and production efficiency. In the following, I will review these results in turn.

2.1.5.1 Phonetic strengthening in dense neighborhoods: Evidence from VOT

VOT (voice onset time) refers to the temporal lag between the release of a stop consonant and the beginning of voicing of the following segment. In English, VOT is a major distinguishing feature for voiceless and voiced stops. Voiced stops (/b/, /d/, /g/) are associated with short VOTs, while voiceless stops (/p/, /t/, /k/) have much longer VOTs.

A preliminary study by Goldinger and Summers (1989) examined the VOTs in high- and low-density words with a word reading task. Speakers were asked to read minimal pairs of CVC words (e.g. *pit* - *bit*) which contrasted in the voicing of the initial stops. Overall, pairs of high-density words exhibited greater differences in VOT than pairs of low-density words, suggesting a hyperarticulatory effect of density. Nevertheless, this study has been criticized for using a paradigm that explicitly draws speakers' attention to the specific contrast under investigation (Wright 1997).

A more recent study by Kilanski (2009) investigated a similar issue, using an improved word-reading paradigm. Kilanski used 12 quadruplets of CVC words, all with voiceless stops ([p], [t], [k]) in the initial position. The four words in a quadruplet were chosen to represent high/low frequency and high/low density, with matching vowel quality. The word list was presented in random orders and subjects were asked to read each word in a carrier phrase ("Say _ again"). Results from 24 subjects showed that there was a weak tendency for high-density words to have longer VOT than low-density words, but the effect was not statistically significant.

Another word-reading experiment by Baese-Berk and Goldrick (2009) studied a slightly different question. Instead of investigating the effects of neighborhood density on VOT, Baese-Berk and Goldrick studied the effects of having particular minimal neighbors. The authors compared the VOTs of voiceless stops in words like *pox*, which had a minimal pair neighbor *box*, and words like *posh*, which did not have such neighbors (**bosh*). Baese-Berk and Goldrick tested with 16 pairs of /p/-initial words, 19 pairs of /t/-initial words and 12 pairs of /k/-initial words. The two words in each pair (*pox-posh*) were matched in frequency, phonotactic probability, phoneme length and vowel type. All critical words were presented in randomized orders, with a large number of fillers. Results showed that words with minimal pair neighbors (*pox*) had longer VOTs than those without such neighbors (*posh*). The trend was significant for each place of articulation, with the greatest effect (20ms difference) being in /p/-initial words. This evidence suggested that the voiceless feature is more accentuated in words with minimal pair neighbors.

Furthermore, Experiment 2 of Baese-Berk and Goldrick (2009) examined the same ef-

fects in context. This experiment involved a speaker and a listener. Both participants were presented with three words on a screen but only the speaker could see which word was selected as the target. The speaker then instructed the listener to select the target by giving commands like "click on the *target word*". Dependent variable is the VOT of the initial stop of the target word in speaker's production. Three conditions of target word presentation were explored. In a Context condition, the target word was presented with its minimal pair neighbor in the closed set (*cod god yell*); in a No Context condition, the target word was presented without its minimal neighbor (*cod lamp yell*); in a No Competitor condition, the target word had no minimal pair neighbor and was presented with two unrelated words (*cop lamp yell*). Thirty six word pairs from Experiment 1 (12 for each place of articulation) were used as the critical stimuli. The results showed that the Context condition produced the longest VOTs (mean = 83ms), followed by the second longest VOTs in the No Context condition (mean = 77ms). Words in the No Competitor condition had the shortest VOTs (mean = 72ms).

Results from Experiment 2 in Baese-Berk and Goldrick's study confirmed that words with minimal pair neighbors were produced with longer VOTs than those with no such neighbors, and further proved that the effect was present in sentence production as well as in isolated word pronunciation. Moreover, the difference between Context and No Context conditions suggested that the VOT difference is even more pronounced when the minimal pair competitor was also present in context.

Overall, the current evidence suggests that the existence of phonological neighbors, especially neighbors that only contrast in the voicing feature of the stop consonant, has a lengthening effect on the VOT of voiceless stops.

2.1.5.2 Phonetic strengthening in dense neighborhoods: Evidence from vowel dispersion

Vowel dispersion is the most frequently-studied variation phenomenon regarding neighborhood characteristics. Various studies have shown that vowels in high-density words are produced with a larger vowel space than those in low-density words (Kilanski 2009; Munson and Solomon 2004; Munson 2007; Watson and Munson 2007; Wright 1997, 2004).

Wright (1997, 2004) analyzed vowel production in 68 CVC words by 10 speakers in a word-reading task. Half of the words had low frequency, high neighborhood density and high neighbor frequency (i.e. "lexically hard" from a perception perspective) and the other half had high frequency, low neighborhood density and neighbor frequency (i.e. "lexically easy"). Segmental context was balanced across conditions. Vowel space expansion in each condition was measured by average vowel dispersion, i.e. the Euclidean distance from individual vowel tokens to the center of the vowel space on the F1-F2 plane (Bradlow, Torretta, and Pisoni 1996). The results showed that vowels in lexically hard words were associated with a more expanded vowel space than those in lexically easy words. In addition, Wright also showed that distance between adjacent vowels was greater in a more expanded vowel space, suggesting that vowels in lexically hard words might be perceptually less confusable with each other

than those in lexically easy words.

However, Wright's study has two caveats. First, frequency covaried with neighborhood characteristics on the word lists, therefore it is possible that the observed effects were only due to word frequency. Second, Wright did not control for vowel duration in the statistical analysis. Since longer vowels tend to be more dispersed, vowel duration is a potential confound for the dispersion results.

Both problems were overcome in a later study by Munson and Solomon (2004). Experiment 1 of Munson and Solomon (2004) replicated Wright's findings with a subset of his stimuli. Morever, Munson and Solomon also analyzed vowel duration in the production data. Interestingly, the analysis revealed an unexpected pattern, that is, vowels in lexically hard words had *shorter* durations (mean = 222 ms) than those in lexically easy words (232 ms), despite their greater dispersion in vowels. Experiment 2 of Munson and Solomon (2004) attempted to separate frequency and (frequency-weighted) neighborhood density. Eighty CVC words were evenly divided into four (high and low frequency × high and low frequency-weighted density) conditions. Analysis of 15 speakers' production confirmed that words from dense neighborhoods were produced with a more expanded vowel space, though the size of the effect was greater for low-frequency words than for high-frequency words. However, different from Experiment 1, no durational difference was found between vowels in high- and low-density words in Experiment 2.

Thus, Munson and Solomon confirmed the effect of density on vowel space expansion, and successfully unconfounded the density effects with frequency and durational variation. In fact, the dissociation of vowel dispersion and duration, replicated in later research by the same author (Munson 2007[6]), is in itself an interesting point. Among other things, it runs counter with the widely accepted notion of "duration-dependent vowel overshoot or undershoot" (Moon and Lindblom 1994), according to which there is a default relation between duration and vowel dispersion simply because given more time, the articulators can reach more extreme positions. Munson and Solomon proposed that shorter vowel durations in lexically hard words might be due to high phonotactic probability. However, this does not explain why vowel production should not be sensitive to phonotactic probability, too.

More recently, Scarborough (in press) examined neighborhood effects on vowel dispersion in context. A set of 48 sentences were used, all containing monosyllabic target words with nasal codas in the final position. The sentences varied orthogonally in (frequency-weighted) neighborhood density (high vs. low) of the target word and contextual predictability (high vs. low). Twelve speakers' reading of the test sentences showed main effects of both variables. High-density words were pronounced with more dispersed vowels than low-density words, and words in less predictable contexts were produced with more dispersed vowels than those in more predictable contexts. However, no interaction of the two variables was found.

Taken together, accumulated evidence regarding vowel dispersion showed that high-density words tend to be produced with a more expanded vowel space, and the effect interacts

[6]But Kilanski 2009 found both lengthening and dispersion in the vowels of high density words.

with word frequency, but not with contextual predictability. A remaining puzzle is why vowel dispersion in high-density words is not accompanied by durational lengthening. I will return to this point in Section 2.1.5.4 below.

2.1.5.3 Phonetic strengthening in dense neighborhoods: Evidence from coarticulation

Scarborough (2004) explored the effects of neighborhood structure on coarticulation. Specifically, two types of coarticulation were investigated: vowel nasalization in monosyllabic words with nasal finals (CVN) and vowel-to-vowel coarticulation in bisyllabic words ending with a vowel /i/ (CVC/i/). Forty CVN words and 40 CVC/i/ words were used as the targets in a modified map task, in which a speaker helped a listener fill out words on an answer sheet. Half of the target words (in both types) were lexically hard and the other half were lexically easy. Speech data were collected from seven speakers, each of whom produced two tokens of each target. For CVN tokens, nasality of the vowel was measured as the relative amplitude of the nasal peak and the first formant (i.e. A1-P0). For CVC/i/ tokens, degree of V-to-/i/ coarticulation was measured by the difference between the F1/F2 of /i/ and that of the canonical /i/ by the same speaker. Scarborough reasoned that since /i/ is the highest and most fronted vowel, coarticulation from any previous vowel would result in a raising in F1 and/or a lowering in F2.

The results showed that overall, lexically hard words were produced with more coarticulation than lexically easy words, as shown in both greater degree of nasality (in CVN words) and raised F1 (in CVC/i/ words). Scarborough interpreted these results as evidence for coarticulation being a form of hyperarticulation, because she assumed that lexically hard words *must* be pronounced with hyperarticulation. However, this line of reasoning is not the only way to interpret the results of the study. The prevailing view in the field considers coarticulation, which is usually indicative of higher degree of gestural overlap, as a feature of casual speech or hypoarticulation (Browman and Goldstein 1992; Lindblom 1990). From this perspective, findings from Scarborough (2004) would suggest that high-density words are *hypoarticulated*, rather than *hyperarticulated*, compared with low-density words. This alternative account has not been considered in Scarborough (2004), but it illuminates the theoretical discussion in the current work, as I will show in the following section.

2.1.5.4 Interpretations of the phonetic strengthening effects of neighborhood density

To sum up, existing studies on neighborhood effects on pronunciation variation all seemed to have found some kind of hyperarticulation in words from dense neighborhoods. The important question is whether it is speaker-centric or listener-oriented considerations that govern variation. With the exception of Baese-Berk and Goldrick (2009), all other studies have argued for a listener-oriented account, which attributes pronunciation variation to (the speaker's) consideration of what will prevent the listener from successful comprehension

(see Section 2.2.5.1 below for more general discussion on the listener-oriented approach). Under this account, high-density (or lexically hard) words are hyperarticulated in order to compensate for their difficulty in perception. However, at least two questions need to be addressed before this account can be considered as convincing.

First of all, it is not clear whether speakers model for perceptual difficulty due to lexical confusability. Bowdle and Wright (1998) did a preliminary study, asking 20 subjects to estimate on a 7-point scale the intelligibility of spoken tokens of 150 CVC words in an imaginary noisy environment. Overall lexically hard words were estimated with lower intelligibility (mean = 3.54) than lexically easy words (mean = 4.31). In the post-experiment survey, more than half of the subjects reported that they considered how many words would sound similar to the stimuli. However, the study did not report how fast the judgments were made, neither did it investigate whether auditory presentation of the stimuli was necessary. Lacking such information, it is hard to judge whether it is possible for normal speakers to do such estimation on a word-by-word basis when planning speech.

More importantly, previous research has shown that neighborhood-conditioned variation exists, regardless of the presence of an addressee in the task (Baese-Berk and Goldrick 2009; Munson and Solomon 2004; Watson and Munson 2007; Wright 1997, 2004), the presence of a lexical competitor in context (Baese-Berk and Goldrick 2009), and predictability of the word from context (Scarborough, in press). Furthermore, previous research by Sommers and colleagues (Sommers and Danielson 1999; Sommers et al. 1997) has found an interaction between density and predictability in speech perception, showing that the inhibitive effects of neighborhood density is reduced in more predictable contexts, but such interaction failed to be observed in pronunciation variation (Scarborough, in press). In other words, speech production is not always sensitive to contextual conditions and listeners' needs. Therefore, even if listener orientation is indeed an underlying motivation for speech modification in lexically hard words, it cannot account for the full spectrum of the variation.

Another question that needs to be addressed is whether variation always results in hyperarticulation in lexically hard words. Durational lengthening and vowel space expansion are both widely acknowledged forms of hyperarticulation (Picheny, Durlach, and Braida 1986), but coarticulation is not. As discussed in the preceding section, coarticulation is usually considered as a form of hypoarticulation, and findings from Scarborough (2004) can be interpreted as evidence for hypoarticulation in high-density words. In addition, the dissociation between duration and vowel dispersion has been found in a number of studies. While vowel dispersion in high-density words was widely observed, a couple of studies (Munson and Solomon 2004; Munson 2007) found that density had no effect on vowel duration. More curiously, in the study by Kilanski (2009), which did find both dispersion and durational lengthening in high-density words, whole word duration was reported to be *shorter* in high-density words than in low-density words. All this evidence suggests that it might be an oversimplified view to attribute all phonetic variation in high-density words to hyperarticulation.

An offline version of the listener-oriented account, the exemplar-based model of speech production, has also been proposed (Munson and Solomon 2004; Pierrehumbert 2002). Un-

der this account, lexically conditioned pronunciation variation is remembered in longterm representations instead of implemented online. Each lexical representation includes highly detailed acoustic information such as VOT length and vowel formant frequencies, which is fleshed out through the user's experiences with perceiving and producing the word in social contexts. Thus, lexically hard words may be better recognized and remembered when they are hyperarticulated, and therefore they tend to be associated with hyperarticulated exemplars in longterm memory for both perception and production. This view will correctly predict no interaction between neighborhood density and contextual predictability, however, as I will show later (§2.2.5.3), it has more general weaknesses in accounting for context-based variations.

In addition to the listener-oriented accounts, an alternative "speaker-oriented" account has been suggested by Baese-Berk and Goldrick (2009), who attributed longer VOTs in potentially more confusable words to higher activation levels in production. Baese-Berk and Goldrick argued that words with minimal pair neighbors are more active when they are produced, compared with similar words with no minimal pair neighbors, and the differences are even larger when the minimal pair neighbors are also present in context. Consequently, words with minimal pair neighbors will be hyperarticulated, and more so when the neighbors are present. This account is in line with the findings of higher production efficiency in high-density words, but the link between higher activity and hyperarticulation is not well-attested in the literature. On the contrary, previous studies seem to suggest that highly activated forms, such as high-frequency and high-predictability words, are more likely to undergo speech reduction rather than hyperarticulation (e.g. Bell, Brenier, Gregory, Girand, and Jurafsky 2009; Jurafsky, Bell, Gregory, and Raymond 2001b), though a theoretical link between activation and reduction is not established, either (see the discussion in §2.2).

To sum up for the current section, neighborhood effects on perception and production efficiency are quite clear: phonological neighbors inhibit word recognition but facilitate lexical production. However, neighborhood effects on pronunciation variation are rather ambiguous. High-density forms have been associated with mostly, though not entirely, hyperarticulated production. Existing accounts have related to both perceptual difficulty (i.e. "listener-oriented" account) or higher activation in production (i.e. "speaker-oriented" account), but both accounts are undermined by problematic arguments. To shed some light on the puzzle, I will turn to a more general literature on pronunciation variation, in search of theoretical links from ease of perception/production to pronunciation variation. In the following section, some major types of within-speaker pronunciation variation will be examined. For each conditioning factor, the effects on perception/production efficiency will also be discussed.

2.2 Within-speaker pronunciation variation

2.2.1 Overview

In this section, I will further the literature review by examining some other types of within-speaker pronunciation variation. The purpose of doing so is to gain insights into the general links between ease of perception/production and pronunciation variation, in order to develop a better understanding of the listener-oriented and speaker-oriented models of speech production. The ultimate goal (see Section 2.3) is to shed light on the underlying mechanisms for neighborhood-conditioned pronunciation variation.

The result of this literature review reveals two recurring patterns: (a) words that are easy to produce tend to be reduced while those that are hard to produce tend to be hyperarticulated; (b) words that are easy to perceive tend to be reduced while those that are hard to perceive tend to be hyperarticulated. The implications of both patterns on models of speech production will be discussed.

It should be noted that this literature review is by no means intended to be an exhaustive list of pronunciation variation phenomena (though a relatively broad coverage will be attempted). Emphasis will be placed on variation phenomena due to lexical conditioning factors that are most similar to neighborhood characteristics. Besides, though variation can occur along a myriad of acoustic dimensions, I will focus the review on variation in duration and vowel dispersion. Not only are these two features most often studied in previous research (including the literature on neighborhood-conditioned variation), they also have the most unambiguous interpretation on the hypo-hyper speech continuum.

In the following, I will briefly introduce the acoustic correlates for pronunciation variation in §2.2.2. Then I will review the major conditioning factors for pronunciation variation in §2.2.3, followed by a discussion on the effects (if any) of the same factors on ease of perception and/or production (§2.2.4). Finally, I will connect the findings reviewed in §2.2.3 and §2.2.4 and reveal two general patterns (§2.2.5).

2.2.2 Phonetic correlates

Pronunciation variation refers to the phenomenon that the same linguistic type (e.g. word, phoneme, etc) can be realized with different acoustic-phonetic features. By saying "different", it is implied that there are some properties of speech that can be reliably compared across tokens. Previous research in the field has investigated variation along various acoustic-phonetic properties, among which duration and degree of vowel dispersion are the most commonly studied. To prepare for the review and discussion in the rest of this section, I will first provide a short description for these two phonetic correlates.

2.2.2.1 Duration

Duration refers to the length in time. It is most often measured for linguistic units such as words and segments. Durational variation can be described in terms of *shortening*

or *lengthening*, though the use of such terms does *not* necessarily entail comparison with the inherent lengths. More often than not, these terms represent the results of comparing with the statistical means. Thus, shortening and lengthening can be two equivalent ways of describing the same variation phenomenon. Interpreting durational variation on the hypo-hyperspeech continuum is straightforward, with shortening associated with hypoarticulation and lengthening associated with hyperarticulation.

2.2.2.2 Degree of vowel dispersion

Degree of vowel dispersion refers to the distance from a vowel token to the center of vowel space. It can be defined either on the articulatory space or on the perceptual space. Depending on how it is defined, degree of vowel dispersion can be associated with different (but correlated) linguistic meanings.

The quality of a vowel is usually characterized by the first two formants (F1 and F2), which roughly correspond to the position of the tongue on the saggital plane of the oral cavity. The F1 dimension indicates the height of the tongue, with higher F1 values associated with lower tongue positions; the F2 dimension indicates the frontness of the tongue, with higher F2 values associated with more fronted tongue positions. Thus, the distance on the F1-F2 plane is indicative of the physical distance that the tongue needs to move for producing two successive vowels. In the meantime, the distance on the F1-F2 plane is also suggestive of perceptual distance, i.e. the amount of differentiation that is perceived by human ears. However, the relationship between actual and perceived frequencies is not linear, as human ears are more sensitive to differences in low frequency ranges than in high frequency ranges (Zwicker 1961, 1975; Zwicker and Terhardt 1980).

Degree of vowel dispersion is often measured as the Euclidean distance from the center of the vowel space on the F1-F2 plane (i.e. $\sqrt{(f1 - f1_{center})^2 + (f2 - f2_{center})^2}$). The greater the distance is, the more dispersed the vowel is. Corresponding to the distinction between articulatory space and perceptual space, degree of vowel dispersion can be associated with two linguistic notions: articulatory effort and perceptual confusability. In the articulatory space, center of vowel space corresponds to the neutral vowel position (i.e. the schwa [ə]), which is produced by opening the jaw without moving the tongue in any dimension. If we assume that the tongue starts with and returns to the neutral position before and after each articulation[7], degree of vowel dispersion is then positively correlated with the amount of tongue movement in vowel production. In the literature on speech reduction (e.g. Jurafsky, Bell, Fosler-Lussier, Girand, and Raymond 1998), degree of vowel dispersion has at times been treated as a binary variable between "reduced" (or "centralized") and "full" .

On the other hand, degree of vowel dispersion has sometimes been measured on perceptully-based scales (e.g. the Bark scale, Zwicker and Terhardt 1980). In this case, center of vowel space simply refers to the geometrical center of the vowel space (encompassed by lines connecting peripheral vowels) and does not have any linguistic meaning. In practice, the center

[7]Note that this is obviously a strong assumption. But the aggregate result of vowel-to-vowel coarticulation with all other vowels might be similar to a schwa.

can be approximated by averaging F1 and F2 of the peripheral vowels, and the choice of peripheral vowels can vary across studies (e.g. Bradlow et al. 1996 used the average F1/F2 of [a], [i] and [o]; Munson and Solomon 2004 used the average F1/F2 of [a], [æ], [ɛ], [ɪ], [i], [o] and [u]). Vowel dispersion defined in this way indicates the radius of the vowel space and therefore is a good estimator for the area of the vowel space (Bradlow et al. 1996). Furthermore, Wright (1997) has shown that dispersion in the perceptual space can also be linked with perceptual confusability, since vowels in a more expanded space are more distant from each other.

In terms of the hypo-hyper continuum, vowel reduction (or centralization) and vowel space contraction are both considered as instances of hypoarticulation, while vowel dispersion and vowel space expansion are both associated with hyperarticulation.

2.2.2.3 Relation between duration and vowel dispersion

As mentioned before, it is widely held that, degree of vowel dispersion (in the articulatory space) is at least partly dependent on duration (Lindblom 1963; Moon and Lindblom 1994). Given more time, the articulators can reach more extreme positions, resulting in more dispersed vowels. Therefore, we would expect to see lengthening to go hand in hand with vowel dispersion and vowel space expansion, and shortening with vowel reduction and vowel space contraction.

However, an important distinction between the two is that variation in vowel quality is more strongly influenced by segmental context. If a vowel is sandwiched between two consonants, the movement of the tongue may not start or end at the neutral position. In this case, shortening may go hand in hand with degree of assimilation, but not necessarily with degree of dispersion (see "Phonetic context" in §2.2.3 for more discussion).

2.2.3 Conditioning factors of pronunciation variation

In the following, I will review a number of conditioning factors for within-speaker pronunciation variation, including frequency, predictability, phonotactic probability, syntactic probability, phonetic context, stress, word class and part of speech, orthography, speech rate, disfluency, repetition, utterance position, noise and features of the listener. The effect of each factor on speech variation will be briefly discussed. Since the purpose of the review is to shed light on neighborhood-conditioned variation, I will emphasize on factors that are most similar to or correlated with neighborhood measures, such as word frequency and phonotactic probability.

2.2.3.1 Effects of frequency on pronunciation variation

Usage frequency is probably the most well-documented conditioning factor for lexically conditioned pronunciation variation. An extremely extensive body of literature has been devoted to this topic. Generally speaking, words that are frequently used tend to be reduced

in speech, as shown in shorter durations (Bell et al. 2009; Gahl 2008; Pluymaekers, Ernestus, and Baayen 2005; Schuchardt 1885; Zipf 1929), reduced vowels (Dinkin 2007; Fidelholtz 1975; Munson 2007; Munson and Solomon 2004; Van Bergem 1995), higher rates of segment deletion (Bybee 2000; Gregory, Raymond, Bell, Fosler-Lussier, and Jurafsky 1999; Hooper 1976; Jurafsky et al. 1998; Jurafsky, Bell, Gregory, and Raymond 2001a; Leslau 1969; Neu 1980) and various types of phonetic assimilation (Ernestus, Lahey, Verhees, and Baayen 2006; Leslau 1969; Patterson and Connine 2000; Phillips 1980, 1983; Rhodes 1992). Evidence comes from various sources, including production experiments, corpus analyses and historical sound changes, for English and other languages in the world (e.g., Ethiopian languages, Leslau 1969; Dutch, Pluymaekers et al. 2005; Taiwan Southern Min, Myers and Li 2009).

However, it would be misleading to think that frequency is *always* a predictor in lenitive variations. No effect of frequency was observed in the tapping of word-final alveolar stops (Gregory et al. 1999), though a strong effect was found on durational shortening and consonant deletion in a related data set. Similarly, frequency was not significant for predicting word-internal t-d deletion (Raymond, Dautricourt, and Hume 2006). This evidence suggests that some lenitive variations are less sensitive (or not sensitive at all) to word frequency than others are.

Furthermore, lenition is not a necessary condition for observing frequency effects, either. There are also sound changes and variations that are not lenitive in nature but also show sensitivity to usage frequency. However, a common feature of these phenomena is a reversed direction of the frequency effect, that is, low-frequency words are more affected than high-frequency words. Hooper (1976) noted that analogical formation started with infrequent words first, since infrequent irregular verbs (e.g. *wept*) shifted to the regular past tense forms (*weeped*) while highly frequent irregular verbs (e.g. *kept*) did not (**keeped*). Phillips (1981, 1984) documented three types of non-lenitive changes: vowel unrounding in Middle English, glide deletion after coronals in Southern American English (e.g. *tune* [tun] and *news* [nuz]) and diatone formation in Modern English (e.g. per'mit vs. 'permit). Phillips showed that none of these sound changes was physiologically motivated and all of them started with low-frequency words first. In light of these observations, Phillips proposed the Frequency Actuation Hypothesis, which postulated that "[p]hysiologically motivated sound changes affect the most frequent words first; other sound changes affect the least frequent words first" (Phillips 1984:336). Though this position might be too strong especially regarding non-physiologically motivated changes, it does point out an important feature of sound changes (and variations) that affect high-frequency items first. High-frequency words are more likely to undergo reductive or assimilative changes, which would result in physiologically more economic productions.

2.2.3.2 Effects of predictability on pronunciation variation

Predictability refers to the likelihood of the occurrence of a word given the surrounding context. There is ample evidence that words in more predictable contexts tend to have more reduced articulation.

Compared with frequency, the assessment of predictability is more variable. The most often cited predictability measures are cloze probability and various collocation-based word predictabilities. Cloze probability is estimated from a cloze test, in which subjects are asked to guess missing words in printed sentences. Collocation-based predictabilities, such as *n-gram* probabilities and other information-theoretic measures, are usually calculated from the frequency counts of word sequences in some corpus. For example, the bigram probability of word W_i given the preceding W_{i-1} is calculated as $P(W_i|W_{i-1}) = C(W_{i-1}W_i)/C(W_{i-1})$, where $C(W_{i-1}W_i)$ is the frequency of the $W_{i-1}W_i$ sequence and $C(W_{i-1})$ is the frequency of W_{i-1}.

No matter how predictability is measured, its effect on speech production is quite robust. Previous research has found a strong tendency for highly predictable words to undergo speech reduction in both elicited speech (e.g. Balota, Boland, and Shields 1989; Lieberman 1963; Shields and Balota 1991) and conversational speech (e.g. Bell, Jurafsky, Fosler-Lussier, Girand, and Gildea 1999; Bell, Jurafsky, Fosler-Lussier, Girand, Gregory, and Gildea 2003; Bell et al. 2009; Gregory et al. 1999; Jurafsky et al. 2001a; Raymond et al. 2006). In this body of research, most of the evidence comes from durational shortening, showing that words in more predictable contexts tend to be shortened. Only a few studies inspected other types of reduction such as vowel centralization (Bell et al. 2003; Jurafsky et al. 2001a)and segment deletion (Jurafsky et al. 2001a; Gregory et al. 1999; Raymond et al. 2006). Compared with durational shortening, other types of speech reduction were not only less studied, but also less sensitive to predictability. Jurafsky et al. (2001a) showed that predictability did not affect the frequency of consonant coda deletion, but it did affect word duration. Similarly, Bell et al. (2003) showed that six of the ten most frequency function words showed predictability effects in word duration, but only three of the words exhibited predictability effects in vowel reduction. These results suggest that, similar to word frequency, predictability has the strongest effects on word duration.

2.2.3.3 Effects of phonotactic probability on pronunciation variation

Phonotactic probability (or phonological pattern frequency) refers to the likelihood of a phonological sequence given the distribution of sounds in the language. In practice, phonotactic probability can be measured based on phonemes, biphones, or syllables. The most widely used phonotactic probability measures include positional phoneme probability (i.e. the probability of specific phonemes in specific positions) and positional biphone probability (i.e. the probability of specific biphone sequences in specific positions) (Vitevitch and Luce 2004).

Munson (2001) showed that high-probability biphone sequences (e.g. /ft/) were produced with shorter durations than low-probability sequences (e.g. /fk/). In addition, Aylett and Turk (2004, 2006) found that both word duration and vowel reduction in connected speech were affected by syllable-to-syllable transitional probability. Syllables that were more predictable from preceding syllables were realized with shorter durations and more reduced vowels.

2.2.3.4 Effects of syntactic probability on pronunciation variation

Syntactic probability refers to the likelihood of the grammatical context. One type of syntactic probability is verb subcategorizational bias. Many English verbs (e.g. *confirm* and *believe*) are subcategorized for either a direct object noun phrase (DO) or a sentential complement (SC). Some verbs in the category (e.g. *confirm*) are more often used with DOs than SCs (i.e. "DO-bias"), while others (e.g. *believe*) have the opposite pattern (i.e. "SC-bias"). Gahl and Garnsey (2004) investigated the relation between verb subcategorizational bias and the pronunciation of words. By analyzing subjects' production in a sentence reading task, Gahl and Garnsey found that verbs and the following NPs were more likely to undergo phonetic reduction in bias-matching contexts (i.e. DO-bias verbs in DO context, SC-bias verbs in SC context) than in the bias-violating contexts.

Another type of syntactic probability, the probability of ditransitive sentence structure, has also been investigated for effects on pronunciation variation. In the English language, a ditransitive sentence can take two surface forms: *V NP NP* and *V NP to NP*. Previous research by Bresnan and colleagues (Bresnan, Cueni, Nikitina, and Baayen 2007; Bresnan and Ford 2010) has shown that the alternation between the two forms is conditioned by the syntax and semantics of the sentence. More recently, Tily, Gahl, Arnon, Snider, Kothari, and Bresnan (2009) found that words occurring in the more probable ditransitive structures were produced with greater fluency and shorter durations.

Taken together, evidence from both studies suggests that words produced within more probable syntactic structures are more likely to be reduced. This tendency is consistent with other frequency- and probability-based effects as reviewed above. Moreover, it suggests that probabilistic effects on pronunciation are not confined to word-to-word or sound-to-sound cooccurrences, as the production system is also sensitive to probabilities of higher-level abstract structures.

2.2.3.5 Effects of phonetic context on pronunciation variation

Depending on what phonemes are in the vicinity, a phonological type can be realized differently and the effect can be shown in both duration and vowel quality. For example, vowels are longer in syllables closed with voiced consonants (e.g. *bead*) than in syllables closed with voiceless consonants (e.g. *beat*) (Klatt 1979; Lehiste 1970). Besides, vowels before syllable-final nasals tend to be nasalized (e.g. *tan*), due to an early start of the articulatory gesture for nasals. Similarly, front vowels are less fronted in syllables closed with [l] and [r] (e.g. *peel* and *hear*), as a result of assimilating to the following retracted approximants.

Most variations in vowel production due to phonetic context can be accounted for by coarticulation. This brings about an important interaction between vowel reduction and phonetic context. As Lindblom (1990) pointed out, though reduced articulatory effort usually leads to vowel centralization, the effect may be countered by coarticulation. For instance, assimilating the [i] vowel in *yeast* to the preceding palatal glide would require less articulatory

effort than reducing the vowel to schwa. Therefore, as mentioned before in §2.2.2, hypo- or hyperarticulation in vowels may not necessarily lead to centralization or dispersion. Instead, hypo- and hyperarticulation may be characterized by degrees of vowel assimilation.

2.2.3.6 Effects of word class and part of speech on pronunciation variation

Word class is another linguistic variable that has been extensively studied in the literature of pronunciation variation. A major distinction is made between content words and function words. Content words (nouns, verbs, adjectives, adverbs, etc) are "open-class" words with semantic meanings, while function words (prepositions, pronouns, auxiliary verbs, etc) are "closed-class" words which are in general meaningless. Since function words are much more frequently used than content words, it follows that function words are more likely to undergo phonetic reduction (Bell et al. 2009; Jurafsky et al. 1998). Furthermore, a couple studies (Bell et al. 2009; Fosler-Lussier and Morgan 1999) have demonstrated that pronunciation variations in function words and content words were predicted by different factors, suggesting that the distinction between function and content words goes beyond a mere frequency effect.

Pronunciation variation also exists among finer syntactic categories. For example, Watson, Breen, and Gibson (2006) showed that nouns were produced with longer durations than verbs, probably because nouns are more likely to be followed by intonational boundaries and therefore more susceptible to phrase-final lengthening.

2.2.3.7 Effects of stress and part of speech on pronunciation variation

Stress has been shown to affect both duration (Beckman 1986; Klatt 1979; Silipo and Greenberg 1999) and segmental realization (Greenberg, Carvey, and Hitchcock 2002). Stressed syllables are associated with longer durations and hyperarticulated vowels, whereas unstressed syllables tend to be shorter and have more reduced vowels. In conversational speech, in addition to lexical stress, certain words in a sentence can receive greater emphasis than others, depending on word category and information structure of the sentence. Generally speaking, function words are usually unstressed, unless they are the foci of the sentence (e.g. ***I*** *didn't eat*).

2.2.3.8 Effects of orthography on pronunciation variation

Even without visual presentation, orthography can affect pronunciation, though evidence is relatively scarce. Warner, Jongman, Sereno, and Kemps (2004) found that Dutch words with the same underlying phonemic content but consistently different spelling (e.g. *kleden* /kledən/ 'to dress/dress (pl.)' vs. *kleedden* /kledən/ 'dressed (pl.)') were pronounced differently. Words with the shorter VC spelling (*kelden*) were produced with slightly shorter durations than the corresponding words with the longer VVCC orthography (*kleedden*). In a different study on English homophonic words, Gahl (2008) found that orthographic reg-

ularity (i.e. grapheme-to-phoneme probability) also affected word duration, as words with more regular spelling forms were pronounced with shorter durations.

2.2.3.9 Effects of speech rate on pronunciation variation

By definition, the rate of speech, which is usually measured in terms of number of syllables per second, will have a great impact on duration and duration-dependent vowel production. Words produced in fast speech tend to have shorter durations and more reduced vowels than those produced in slow speech. The predictions have been confirmed in various corpus studies (Bell et al. 2003, 2009; Fosler-Lussier and Morgan 1999; Gregory et al. 1999; Jurafsky et al. 2001a).

2.2.3.10 Effects of disfluency on pronunciation variation

Speech production is a complicated process and the flow of speech is not always fluent. Disfluency in production can take various forms, such as unfilled pauses, filled pauses (e.g. *um* and *uh*), word repetitions (*that that*), cutoff words (e.g. *th- they*), repairs (e.g. *any health cover - any health insurance*), etc. Not surprisingly, spontaneous speech is more susceptible to disfluency compared with other speech styles. As shown in Shriberg (2001), the rate of disfluency is about is 0.06 per word (i.e. 6 disfluencies per 100 words) in spontaneous speech, but only 0.008 in the more constrained human-computer dialogue.

The most well-researched type of disfluency is associated with hesitation (e.g. filled and unfilled pauses, repetition). Hesitation-related disfluency generally signals problems in speech planning and often contains a prospective source in upcoming words. The phonetic consequence of such disfluency is often shown on words that precede the point of disfluency, resulting in longer durations (Duez 1993; Eklund and Shriberg 1998; Lickley 1994; O'Shaughnessy 1992; Shriberg 1999) and fuller pronunciations (Fox Tree and Clark 1997). A well-known study by Fox Tree and Clark (1997) investigated how the pronunciation of the article *the* alternated between [ðiː] with a tense vowel and [ðə] with a reduced vowel. Their results from a corpus revealed that most (81%) of the tense vowel pronunciations were followed by suspension in speech whereas a much smaller proportion (9%) of the reduced vowel pronunciations were followed by suspension. In other words, pronouncing *the* as [ðiː] was strongly correlated with a cease of fluent speech and upcoming hesitation. The same trend has been observed in other function words (e.g. *to* and *a*) with similar pronunciation alternations (Bell et al. 1999, 2003; Shriberg 1999).

Words that follow hesitation-related disfluency have also been associated with longer durations and fuller vowel forms, but to a much lesser degree compared with words preceding disfluency (Bell et al. 2003; Shriberg 2001).

2.2.3.11 Effects of discourse repetition on pronunciation variation

Words that have occurred in previous context tend to be shortened. Fowler and Housum (1987) examined utterances in monologs, which contained either first or second mentions

of the critical words, and found that the second mentions were significantly shorter than the first ones. The trend was confirmed by other experiments (Bard, Anderson, Aylett, Doherty-Sneddon, and Newlands 2000; Fowler 1988; Hawkins 2003) and corpus studies (e.g. Bell et al. 2009; Gregory et al. 1999). Moreover, Bell et al. (2009) compared several ways of coding repetition, using both a binary coding ("first mention" vs. "repeated mention") and continuous measures (e.g. raw and log number of repetitions). The results showed that the binary variable was the most effective for predicting repetition-based durational shortening, followed by log number of repetitions. Raw number of repetitions did not reach significance. This evidence suggested that the greatest distinction was between the first and second mentions of a word. Further shortening in the mentions subsequent to the second one was minimal.

2.2.3.12 Effects of utterance position on pronunciation variation

Words at the beginning and end of an utterance are subject to phrase-boundary intonation. A major phonetic consequence is phrase-initial and phrase-final lengthening (Bell et al. 2003; Byrd, Kaun, Narayanan, and Saltzman 2000; Crystal and House 1990; Klatt 1975, 1979; Ladd and Campbell 1991). Another pattern that is associated with utterance production is the overall declination of articulatory forces, as reflected in stronger initial segments (e.g. Fougeron and Keating 1997; Byrd et al. 2000) and weaker final segments (e.g. Browman and Goldstein 1992) (though evidence for final strengthening of vowels has also been documented in Fougeron and Keating 1997, which may be due to final lengthening).

2.2.3.13 Effects of the environment and the listener on pronunciation variation

Features of the environment and the listener can also affect pronunciation variation. In general, speakers talk louder, produce more hyperartciculated segments and employ wider pitch ranges in noisy environment ("Lombard speech", Lau 2008; Lombard 1911; Summers, Pisoni, Bernacki, Pedlow, and Stokes 1988), and when the listener is a child (Baldwin and Baldwin 1973; Broen 1972; Remick 1971; Sachs, Brown, and Salerno 1976; Snow and Ferguson 1977) or a nonnative speaker (Long 1981).

This type of variation is often attributed to a general distinction in speaking style between *clear speech* and *casual speech*. It has been shown that speakers adopt a more careful speech style when reading citation forms and/or explicitly instructed to speak clearly (Cutler and Butterfield 1991; Lively, Pisoni, Summers, and Bernacki 1993; Moon and Lindblom 1994). Pronunciation in clear speech is in general associated with slower speaking rates, longer durations and vowel space expansion (Bond and Moore 1994; Payton, Uchanski, and Braida 1994; Picheny et al. 1986; Uchanski, Choi, Braida, Reed, and Durlach 1996), though more recent evidence from research by Krause and Braida (Krause and Braida 2002, 2004) showed that it is also possible to produce fast *and* clear speech. Compared to other conditioning factors reviewed above, speech style has a more global effect on pronunciation variation which is not limited to certain words or segments.

2.2.3.14 Summary of conditioning factors of pronunciation variation

In the above, I have reviewed some major types of within-speaker pronunciation variation (see Table 2.2 for a summary). As shown in Table 2.2, most of the conditioning factors can be associated with reduction (shortening, vowel reduction, segmental deletion) or hyperarticulation (lengthening, vowel dispersion, segmental preservation)[8]. A few factors are also associated with assimilation, which, like reduction, is featured by reduced articulatory effort (Browman and Goldstein 1992; Lindblom 1990). Thus, it suffices to say that a general pattern in Table 2.2 is that pronunciation variation reflects changes in the strength of articulation. One may argue that this is an automatic result of my decision to emphasize on the literature regarding variation in duration and vowel dispersion, because these two phonetic features are most ready to be interpreted on the hypo-hyper speech continuum. However, as mentioned before, a major reason for choosing these two features was because of their extremely wide spread in the literature. Therefore, it is not unwarranted to consider the pattern in Table 2.2 as a general tendency for within-speaker pronunciation variation.

A secondary pattern in Table 2.2 is that frequency and various probabilistic measures (predictability, phonotactic probability, syntactic probability) behave in highly similar ways. An extremely robust tendency is that high-frequency/probability words are more likely to undergo reduction. If we consider word frequency as an *a priori* probability, i.e. the probability of a word without context, then frequency and other probabilities can be united under the notion of *word probabilities*, which describe the overall likelihoods of a word form. The general relationship between word probabilities and speech reduction has been widely attested (Aylett and Turk 2004, 2006; Jurafsky et al. 2001b), and has motivated a number of theoretical accounts (the Probabilistic Reduction Hypothesis, Jurafsky et al. 2001b; the speech efficiency hypothesis, Van Son and Pols 2003; the Smooth Signal Redundancy Hypothesis, Aylett and Turk 2004; the Universal Information Density theory, Levy and Jaeger 2007). Most of these theories are from an information-theoretic point of view, linking probability with information load and entropy of the word. However, the pattern captured in Phillips (1984)'s Frequency Actuation Hypothesis suggests that the link should also exist in speech production; otherwise we would not expect to see the direction of the frequency effects being dependent on whether the variation is physiologically motivated or not.

Another pattern revealed in the literature review (but not shown in Table 2.2) is the independence between duration and vowel dispersion. Generally speaking, shortening (lengthening) and vowel reduction (dispersion) go hand in hand, and there is clearly a physiological reason for such correlations (Lindblom 1963; Moon and Lindblom 1994). However, several patterns of dissociation between the two have been documented, including shortening beyond vowel reduction (Bell et al. 2003; Jurafsky et al. 2001b) and vowel dispersion without lengthening (Krause and Braida 2002, 2004). These findings remind us of the dissociative pattern observed in Munson and Solomon (2004). Importantly, what has *not* been observed so far is when duration and vowel dispersion go in opposite directions, resulting in combinations in

[8] As noted before, some anti-frequency effects cannot be characterized by any of these terms. But these case are clearly in the minority and possibly motivated by different mechanisms.

Category	Factor	Effect on pronunciation variation
Linguistic	Frequency	Reduction; assimilation
	Predictability	Reduction
	Phonotactic probability	Reduction
	Syntactic probability	Reduction
	Phonetic context	Phonetically-conditioned variation, mostly assimilation.
	Word class	More reduction in function words than in content words
	Part of speech	Verbs are shorter than nouns
	Orthographic regularity	Reduction
	Stress	Hyperarticulation
Extra-linguistic	Speech rate	Reduction
	Disfluency	Hyperarticulaton
	Repetition	Reduction
	Utterance position	Lengthening at utterance boundaries Initial strengthening, final weakening
	Noise	Hyperarticulation
	Nonnative addressee	Hyperarticulation
	Child addressee	Hyperarticulation

Table 2.2: Conditioning factors for pronunciation variation and the associated effects. Here the term "reduction" is used as a cover term for durational shortening, vowel reduction and segmental deletion, and the term "hyperarticulation" is used as a cover term for durational lengthening, vowel dispersion and segmental preservation. If the effect is simply listed as "hyperarticulation" or "reduction", it should be understood that the effect is associated with an increase in the conditioning factor if the factor is continuous (e.g. frequency), or with the TRUE level of the factor if it is binary (e.g. stress and disfluency). It should also be noted that empty cells in the table do not necessarily indicate null effects. In some cases (e.g. effects of syntactic probability and orthographic regularity on production efficiency), the empty cells represent gaps in the literature (or more likely, gaps in the literature review).

the forms of shortening+vowel dispersion and lengthening+vowel reduction.

In the next part of literature review, I will turn to a different theoretical question, that is, what are the effects of the same conditioning factors on speech perception and production efficiency. The goal of this literature review is to link the variations reviewed above with ease/difficulty of perception and production.

2.2.4 Relationship with ease of perception/production

The previous section reviewed some major types of within-speaker pronunciation variation. In this part of the literature review, I will discuss the perceptual and production basis of the variations. Since the ultimate goal is to gain insights into the listener-oriented and speaker-oriented models of speech, I will emphasize factors that affect ease of auditory perception and production.

Several factors reviewed in the previous section will not be included in the following discussion. These factors can be divided into three categories. The first category includes stress and speech rate, both of which have default relationships with reduction and hyperarticulation. Phonetic reduction (especially durational shortening) is a defining feature of unstressed pronunciation and fast speech, therefore it will be a vain attempt to seek perceptual and production basis for such variations. The second category consists of conditioning factors that lead to pronunciation variation via mechanisms outside the realms of speech perception and production. Such factors include word class, part of speech and utterance position. Variations due to these factors are mediated by prosody (e.g. nouns are longer than verbs because they are more often at phrase-final positions) and therefore are not immediately related with ease of perception or production. The last category consists of disfluency. Despite some disagreement in the literature[9], disfluency is largely considered to be a consequence of production difficulty. In other words, disfluency does not facilitate or impede production - it reflects the difficulty in planning and/or articulation, which may be attributed to other linguistic or nonlinguistic factors.

The rest of the conditioning factors (frequency, predictability, phonotactic probability, syntactic probability, orthography, disfluency, discourse repetition, features of the environment and the listener) have all been associated with facilitation or inhibition in auditory perception and/or speech production. In the rest of this section, I will review the empirical evidence for these effects.

2.2.4.1 Effects of frequency on perception and production efficiency

High-frequency words are easier to recognize than low-frequency words (e.g. Marslen-Wilson 1987, 1990; Luce and Pisoni 1998). As reviewed in §2.1.3, the facilitative effects of

[9]Some research on disfluency (e.g. Clark and Fox Tree 2002; Ferreira and Bailey 2004; Fox Tree and Clark 1997) has suggested that speakers may use disfluency to signal various information to the listener, which may in fact facilitate speech comprehension.

frequency have been obtained from a variety of experimental paradigms (perceptual identification, auditory lexical decision, word naming, etc), in terms of both recognition latency and accuracy. Most current models of speech recognition (Cohort, Shortlist, TRACE, NAM, PARSYN) agree that high-frequency words have an advantage in processing, and have considered higher resting activation levels (Dahan, Magnuson, and Tanenhaus 2001; McClelland and Rumelhart 1981; Marslen-Wilson 1990), stronger connections (Dahan et al. 2001; Gaskell and Marslen-Wilson 1997; MacKay 1982, 1988; Plaut, McClelland, Seidenberg, and Patterson 1996) and biases in selection rules (Dahan et al. 2001; Goldinger et al. 1989; Luce 1986; Luce and Pisoni 1998) as the underlying mechanisms for such advantage. All the proposed mechanisms are mutually compatible (Dahan et al. 2001), and so far no evidence has been presented to exclude any possibility.

Frequency also facilitates word production. A voluminous body of experimental records (Balota and Chumbley 1985; Carroll and White 1973; Dell 1990; Griffin and Bock 1998; Jescheniak and Levelt 1994; Oldfield and Wingfield 1965) has demonstrated that high-frequency words are produced with shorter latencies and higher accuracy compared with low-frequency words. Evidence comes from picture naming (Carroll and White 1973; Griffin and Bock 1998; Jescheniak and Levelt 1994; Oldfield and Wingfield 1965; Wingfield 1967), printed word naming (Balota and Chumbley 1985; Forster and Chambers 1973), printed lexical decision (Forster and Chambers 1973), semantic and grammatical categorization (Bonin and Fayol 2002; Jescheniak and Levelt 1994), and naturalistic and elicited speech errors (Dell 1990; Vitevitch and Sommers 2003).

The exact locus of the frequency effect on production efficiency is still under debate. Presumably all stages of speech production (conceptual preparation, lemma selection, retrieval of phonological forms, articulation) can exhibit some usage frequency effects, but the type of frequency that each stage is sensitive to may be different. For example, conceptual preparation is probably sensitive to the frequency of concepts being invoked, which is different from the frequency of a word being used, due to the n-to-n correspondence between concepts and words. Previous research (Dell 1990; Jescheniak and Levelt 1994) has shown that a major source of the word frequency effect lies in lexical access, especially the retrieval of phonological forms. It has been demonstrated that frequency effects persist with the same magnitude after controlling for the complexity of conceptualization (Almeida, Knobel, Finkbeiner, and Caramazza 2007; Jescheniak and Levelt 1994; Wingfield 1967), but quickly dissipate (Jescheniak and Levelt 1994; Munson 2007) or at least significantly reduce in magnitude (Balota and Chumbley 1985) in the articulation stage. Furthermore, both Dell (1990) and Jescheniak and Levelt (1994) showed that low-frequency words with high-frequency homophones patterned with high-frequency words in terms of vulnerability to speech errors and production latencies. This type of "frequency-inheritance" among homophonic words suggests that frequency is a property of phonological forms. However, more recent research by Gahl (2008; see also Bonin and Fayol 2002) challenged this position by showing that the production of homophonic words is not completely identical – instead it is conditioned by the frequency of individual homophonic words. This evidence suggests that speech production is also conditioned by lemma frequency.

Overall, it suffices to say that high-frequency words are easier to produce than low-frequency words and a major source of this effect is in lexical access.

2.2.4.2 Effects of predictability on perception and production efficiency

Facilitative effects on perception have also been observed with contextual predictability. Words are recognized faster and more successfully when preceded by semantically related primes (e.g. Moss, Ostrin, Tyler, and Marslen-Wilson 1995). Similar facilitative effects have been obtained with a sentence context which may or may not include any directly related words. Evidence of this type comes from a variety of experimental paradigms, including crossmodal priming (Zwitserlood 1989), word identification (Boothroyd and Nittrouer 1988; Craig, Kim, Rhyner, and Chirillo 1993; Sommers and Danielson 1999), speech shadowing (Marslen-Wilson 1985; Marslen-Wilson and Welsh 1978), mispronunciation detection (Cole and Perfetti 1980), phoneme and word monitoring (Marslen-Wilson and Tyler 1980) and gating (Grosjean and Fitzler 1984; Zwitserlood 1989).

In addition to the main effect of facilitation, predictability also interacts with frequency, resulting in a reduction of the frequency effect in high-predictability contexts. Grosjean and Fitzler (1984) noted in their gating study that the gates for successful identification were earlier in high-frequency words than low-frequency words, but the frequency effect was reduced when the words were more predictable from context, and almost went away in the most constraining contexts.

The effect of contextual predictability in auditory perception has mostly been attributed to lexical activation levels. During word recognition in context, activation of lexical candidates is affected by both higher-level contextual constraint and lower-level acoustic-phonetic information. As Zwitserlood (1989) stated, "context positively affected the activation level of contextually appropriate candidates at a point in time where the sensory information was insufficiently constraining to distinguish between appropriate and inappropriate candidates" (p150).

By comparison, there have been relatively fewer studies on the effect of contextual predictability on production efficiency. Balota et al. (1989) showed that semantically related word pairs had shorter production latencies than unrelated pairs. Griffin and Bock (1998) extended the findings to word production in sentence contexts. With a contextualized picture naming experiment, in which subjects were presented with unfinished sentences prior to the picture targets, Griffin and Bock showed that naming latencies were negatively correlated with the degree of contextual constraint provided by the preceding sentences. The shortest naming latencies were produced in the most constraining contexts. Furthermore, similar to Grosjean and Fitzler (1984), Griffin and Bock also found an interaction of frequency and predictability. Differences between high- and low-frequency words were smaller in more predictable contexts, and almost reduced to zero in the most constraining contexts.

The locus of the predictability effect in production is generally agreed to be in lexical selection (e.g. Griffin and Bock 1998; Roelofs 1992; Stemberger 1985). In highly predictable contexts, fewer lemmas are consistent with the contextual constraint, therefore, lexical se-

lection proceeds with a smaller set of competing lemmas, which results in a faster selection process.

To sum up, contextual predictability facilitates both speech perception and production. In both cases, the locus of the effect is in lexical access. What's more, an interaction with frequency effects has been observed in both perception and production. Differences between high- and low-frequency words are less obvious in more predictable contexts, and can be completely eliminated in highly constraining contexts.

2.2.4.3 Effects of phonotactic probability on perception and production efficiency

As reviewed in §2.1.3 and §2.1.4, phonotactic probability facilitates both speech perception and production. Spoken stimuli (both word and nonword) of higher phonotactic probability are recognized faster than those of lower phonotactic probability (Luce and Large 2001; Vitevitch and Luce 1998, 1999). The facilitative nature of phonotactic probability in speech perception puts it in direct contrast with neighborhood density, which *inhibits* auditory word recognition (Luce and Pisoni 1998). As reviewed in §2.1.3, previous research (Vitevitch and Luce 1998, 1999) has successfully separated these two variables in speech perception by showing that neighborhood density inhibits processing at the lexical level whereas phonotactic probability facilitates processing at the sublexical level.

In terms of production, high-probability phonological sequences also have shorter production latencies than low-probability sequences, and the effects exists for both words and nonwords (Munson 2001; Vitevitch et al. 2004). The locus of the effect in production has been attributed to phonological encoding in production (Munson 2001).

2.2.4.4 Effects of syntactic probability on perception and production efficiency

Sentences with more probable structures are easier to process. Specifically, in terms of subcategorization bias of the verb, sentences with bias-matching structures were processed more quickly and smoothly than those with bias-violating structures (Clifton et al. 1984; Tanenhaus, Stowe, and Carlson 1985; Trueswell, Tanenhaus, and Kello 1993).

By contrast, there has been no equivalent studies on the effect of syntactic probability on production efficiency, though there exist studies on pronunciation variation (e.g. Gahl and Garnsey 2004). Given what we know so far about the general probabilistic effects on production efficiency, it seems reasonable to hypothesize that sentences with more probable structures will also be easier to produce.

2.2.4.5 Effects of orthography on perception and production efficiency

Current evidence suggests that orthographic regularity may facilitate both perception and production, but the evidence is relatively weak, and the presence of the effects may be subject to the nature of the task. The effects of orthographic length, on the other hand, are

largely unknown, though there is ample evidence of facilitation in shorter words from the visual domain (see New, Ferrand, Pallier, and Brysbaert 2006 and references therein).

Ziegler and colleagues (Ziegler and Ferrand 1998; Ziegler, Ferrand, and Montant 2004) demonstrated that the degree of spelling consistency predicted performance on auditory recognition tasks. Words with rimes that can only be spelled in one way (e.g. /-ʌk/ can only be spelled as "-uck") were responded to faster than those with rimes that can be spelled in multiple ways (e.g. /-aɪn/ can be spelled as either "-ine" or "-ign") (Ziegler and Ferrand 1998). Furthermore, among words with multiple possible spellings, those with dominant spelling forms (e.g. "-ine" is the dominant spelling for /-aɪn/, as shown in *wine, pine, mine, nine, fine, dine, line*, etc) were easier to recognize than those with subdominant spelling forms (e.g. "-ign" only exists in *sign* and a couple other words) (Ziegler et al. 2004). Taken together, stronger grapheme-to-phoneme correspondence, or higher orthographic regularity, facilitates auditory word perception. In addition, Ziegler et al.'s work also showed that the spelling effects were stronger in tasks that require lexical activation (e.g. lexical decision) than in those that don't (e.g. rime detection and auditory naming).

Orthography may also affect production efficiency (Alario, Perre, Castel, and Ziegler 2007; Damian and Bowers 2003; Lupker 1982; Roelofs 2006; Tanenhaus, Flanigan, and Seidenberg 1980; Ventura, Kolinsky, Querido, Fernandes, and Morais 2007; Wheeldon and Monsell 1992). Among others, Damian and Bowers (2003) showed that words sharing both initial phonology and orthography (e.g. *camel* and *coffee*) facilitated each other in production, but words that only shared phonology (e.g. *kennel* and *coffee*) did not. These results suggest that the locus of the spelling effect is at the intersection of phonology and orthography. Though the study did not investigate orthographic regularity *per se*, the results lead us to expect that words with high grapheme-to-phoneme probabilities will be facilitated in production because they receive more support from words that share the orthography.

However, what is still under debate is whether the spelling effects in production rely on the nature of the task. While Damian and Bowers (2003) showed that the effect persisted even when the stimuli were presented auditorily, Roelofs (2006) (see also Alario et al. 2007) found the effect only present in oral reading, where orthography was presented, but not in object naming or word generation. Thus, it is not clear whether orthography would also affect conversational speech production.

2.2.4.6 Effects of discourse repetition on perception and production efficiency

Repetition facilitates both perception and production. Evidence is abundant in the visual domain (e.g. Forster and Davis 1984), but less so in the auditory domain. A number of auditory recognition studies (Radeau et al. 1989; Slowiaczek and Pisoni 1986) found that recognition was facilitated when preceded by identical primes, and the effects were much more robust than partially overlapping primes.

The repetition effect on production efficiency is extremely robust and well-documented. Various naming studies showed that target word production is facilitated if the same word has somehow been recently produced by the subject, either in the same naming paradigm

(Barry, Hirsh, Johnston, and Williams 2001; Biggs and Marmurek 1990; Durso and Johnson 1979), or as a response to a definition (Monsell, Matthews, and Miller 1992; Wheeldon and Monsell 1992) or a translation task (Francis, Augustini, and Sáenz 2003; Francis and Sáenz 2007), or simply in a read-aloud task before the naming experiment (Barry et al. 2001; Wheeldon and Monsell 1992). The repetition effects are also robust enough to sustain long intervals (as long as 5 min) between two productions (Durso and Johnson 1979) and a large number (> 100) of intervening trials (Wheeldon and Monsell 1992).

The site of the repetition effect (at least in production efficiency) has been argued to be at the intersection of lexical meaning and phonological form (Monsell et al. 1992; Wheeldon and Monsell 1992). Evidence comes from cross-linguistic tasks and homophonic studies. Semantically equivalent words from different languages (which hence differ in phonological forms) do not facilitate each other in bilingual speakers' production (Monsell et al. 1992). On the other hand, homophonic words which share phonology but differ in meaning also fail to facilitate each other in production (Wheeldon and Monsell 1992). Thus, it is strongly suggested that the locus of the repetition effect is in the access of both lemma and lexeme. Another type of supporting evidence comes from the interaction between repetition and frequency. It has been observed that differences between high- and low-frequency forms diminish over repetition (Bartram 1973, 1974; Griffin and Bock 1998; Oldfield and Wingfield 1965; Scarborough, Cortese, and Scarborough 1977; but see Levelt, Praamstra, Meyer, Helenius, and Salmelin 1998), and the repetition effect also attenuates among high-frequency words (Forster and Davis 1984; Howard and Burt 2010; La Heij, Puerta-Melguizo, van Oostrum, and Starreveld 1999; Wheeldon and Monsell 1992). These interactions suggest that the locus of the repetition effect overlaps with that of the frequency effect, which is mainly in lexical access.

There are two possible accounts for the observed repetition effects. The first one connects repetition with recency and ascribes the effects to residual activation after prior exposure. This account is in line with previous findings of recently encountered words affecting subsequent speech production (Griffin 2002). However, a corollary of this position is that repetition effects should be transient, as residual activation will quickly dissipate. Apparently this is not supported by the empirical records which found long-lasting repetition effects (Durso and Johnson 1979; Wheeldon and Monsell 1992; though Ferrand 1996 presented evidence for short-lived repetition effect in printed word naming).

An alternative account connects repetition with frequency and attributes the effects to changes in long-term representations and interconnections in the lexicon. Every time a word is repeated, its resting activation level and connection strengths may be enhanced, which will give rise to a frequency effect after constant repetition. This account will correctly predict long-lasting repetition effects, but cannot explain the recency effect.

Since the two mechanisms are mutually compatible, it seems most likely that both of them are at work (Forster and Davis 1984). Put in another way, after each repetition, both instantaneous and long-term activation patterns of the word will be updated, which will affect lexical access in future processing.

2.2.4.7 Effects of the environment and the listener on perception and production efficiency

Environmental noise inhibits both speech perception and production. The noise effect on speech perception is obvious (Payton et al. 1994), and has been widely evidenced by human subjects' degraded performance in perceptual tasks with noise-covered stimuli and/or adverse listening conditions. Background noise also makes speech production more difficult, by way of obstructing auditory feedback. In normal speech production, the speaker is constantly monitoring the acoustic output, in order to correct errors and adjust subsequent production. When auditory feedback is obstructed by noise, speech monitoring becomes difficult, which may in turn impede speech production. Ladefoged (1967: 163-165) reported an informal study which examined speech production under noise conditions. It was noted that when auditory feedback was completely eliminated by a loud masking noise, the subjects' speech production became "very disorganized".

Some features of the listener affect the perception efficiency of the listener. Compared with native adult listeners, both nonnative listeners and child listeners have less efficient perception systems. Previous research has attributed the difficulty in L2 perception to interference from L1 phonology (Weber and Cutler 2004) and an overall lower sensitivity to fine-grained acoustic-phonetic detail in L2 stimuli (Bradlow and Bent 2002; Bradlow and Pisoni 1999). The perceptual difficulty in children is mostly due to the underdevelopment of their language system, especially regarding vocabulary size and the degree of phonetic specification of lexical representations.

There is no evidence that facing a nonnative or child listener will affect the production efficiency in the speaker in any way, and there is also no reason to expect such effects to be true.

2.2.4.8 Relating pronunciation variation with ease of perception and production

In the above, I have shown that many conditioning factors for pronunciation variation also affect speech perception and production efficiency (see Table 2.3 for a summary). As shown in Table 2.3, words of high probabilities (frequency, predictability, phonotactic and syntactic probability) and high orthographic regularity as well as those repeated from previous context are easier to recognize. By contrast, words produced in noisy environment are harder to perceive, and perception is in general more difficult by nonnative listeners and child listeners. In terms of production, words of high probabilities (except for syntactic probability, for which there is no empirical evidence) and those repeated from previous context are produced more easily and successfully. Noisy environment in general also impedes speech production.

Table 2.3 also shows the effects of these factors on pronunciation variation from previous discussion (repeated from Table 2.2). If we compare across columns, two patterns can be observed. The degree of articulatory strength, as reflected on the hypo-hyperspeech continuum, can be associated with either ease of perception or ease of production. On one

Factor	**Effects on listener's perception**	**Effects on production efficiency**	**Effects on pronunciation variation**
Frequency	Facilitation	Facilitation	Reduction and assimilation
Predictability	Facilitation	Facilitation	Reduction
Phonotactic probability	Facilitation	Facilitation	Reduction
Syntactic probability	Facilitation	–	Reduction
Orthographic regularity	Facilitation	–	Reduction
Repetition	Facilitation	Facilitation	Reduction
Nonnative listener	Inhibition	–	Hyperarticulation
Child listener	Inhibition	–	Hyperarticulation
Noise	Inhibition	Inhibition	Hyperarticulation

Table 2.3: Conditioning factors for pronunciation variation and their effects on perception, production efficiency and pronunciation variation. Here the term "reduction" is used as a cover term for durational shortening, vowel reduction and segmental deletion, and the term "hyperarticulation" is used as a cover term for durational lengthening, vowel dispersion and segmental preservation. All listed effects are associated with increase in the conditioning factor if the factor is continuous, or with the TRUE level of the factor if it is binary. It should be noted that empty cells in the table do not necessarily indicate null effects. In some cases (e.g. effects of syntactic probability and orthographic regularity on production efficiency), the empty cells represent a gap in the literature (or more likely, a gap in the literature review).

hand, when perception is facilitated (as in the case of high frequency, predictability, etc), the pronunciation of the word tends to be reduced, and when perception is obstructed (as in the case of background noise, and non-native and child listeners), the word tends to be hyperarticulated. On the other hand, when production is facilitated (as in the case of high frequency, predictability, etc), word production tends to be reduced or assimilated, and when production is inhibited (as in the case of background noise), the word tends to be hyperarticulated. That is to say, both ease of perception and ease of production are positively correlated with the rate (or likelihood) of speech reduction. These two associative patterns are largely confluent, as the conditioning factors in Table 2.3 all have the same effects on perception and production (if they affect both). However, the two patterns have motivated differing accounts for the underlying mechanisms of pronunciation variation. In the following, I will explain the two accounts in turn.

2.2.5 Underlying mechanisms

2.2.5.1 Listener-oriented account

The connection between ease/difficulty of perception and pronunciation variation can be explained by listener-oriented speech tailoring. It is widely assumed that speakers design their utterances in order to suit the needs of their listeners (that are known to the speakers) (Clark and Marshall 1981; Lindblom 1990). Tailoring can occur in different levels, but most current evidence comes from tailoring in the form of referring expressions and the clarity of speech. For the purpose of the current discussion, I will emphasize more on the latter type of tailoring.

A highly influential theory, the Hypo- and Hyperspeech (H&H) theory by Lindblom (Lindblom 1990), proposes that speakers devote more effort to pronunciation when they anticipate perceptual difficulty by their listeners. The proposed strategy serves two purposes: First, it ensures speech intelligibility; second, it optimizes the expenditure of articulatory effort. Following this claim, one would expect that only words that are hard to recognize will be hyperarticulated and those that are easy to recognize should be reduced and assimilated.

The strongest evidence for the H&H theory comes from speech style change in response to conversation situations. As mentioned before, speakers tend to deliver clearer speech (a) in presence of background noise, (b) when knowing that the listener has suboptimal language proficiency, and (c) when instructed to do so. In the latter two cases, the adoption of a clear speech style, or global hyperarticulation, is clearly motivated by the perceived needs of the listener. In the case of background noise, although it has been shown that the Lombard effect exists even when there is no listener around (Summers et al. 1988), suggesting that the effect is physiological (Junqua 1996), a recent study by Lau (2008) found that the effect also exists when *only* the listener, but not the speaker, is exposed to noise, which indicates that listener adaptation is at least part of the motivation for hyperarticulation under noise. Furthermore, previous research has shown that this type of global speech modification is indeed effective, as clear speech in general has higher intelligibility than casual speech (Bradlow et al. 1996;

Bond and Moore 1994; Payton et al. 1994; Picheny, Durlach, and Braida 1985).

It has also been claimed that change in speech style can occur on a finer scale, in response to anticipated perceptual difficulty in individual words. Variation in pronunciation clarity with regards to contextual predictability (Hunnicut 1985; Lieberman 1963) and repetition (Fowler and Housum 1987) is often attributed to how expected the information is to the listener (as modeled by the speaker). More expected information will have less articulated forms while less expected information will have more emphasized forms. However, Bard and colleagues (Bard et al. 2000; Bard and Aylett 2005) have shown that reduction in repeated mentions exists, even when speakers can easily infer, either from previous absence of the listener or explicit negative feedbacks, that the information is *not* expected by the listener. That is to say, pronunciation variation at the word level cannot be fully attributed to listener orientation.

More generally, the claim of word-based intelligibility tailoring is undermined by two theoretical issues. The first one concerns the computational feasibility of tracking perceptual difficulty online. The second issue concerns the time efficiency of word-based intelligibility tailoring. In the following, I will elaborate on both issues.

Tracking perceptual difficulty online entails that among other things, the speaker needs to maintain a model of the listener's knowledge and beliefs which also includes what the listener knows that the speaker knows (i.e. mutual knowledge) (Clark and Marshall 1981). As conversation unfolds, this task is increasingly complicated and may eventually become insuperable. A number of alternatives have been proposed, such as the *copresence default hypothesis* (Bard et al. 2000) and the *monitoring and adjustment hypothesis* (Horton and Keysar 1996), both of which posit that instead of constantly updating the listener model, speakers may default to their own knowledge systems and only make adjustments and repairs when necessary. Doing so will lessen the computational burden of online monitoring and recalculation, however, it inevitably compromises the accuracy of the listener model.

In addition, even if we assume that the listener model is always correct, there is still a problem of time efficiency. If intelligibility tailoring is to be implemented on a word-to-word basis, the production system will need to consult the listener model at the production of each word. The question is whether the intervals between consecutive word productions are long enough for both drawing inference from the listener model and designing subsequent articulation. As mentioned above, Bard and Aylett (2005) showed that when speakers can clearly infer listener's unfamiliarity with novel objects, their pronunciation of the object names still exhibits reduction as compared to first mentions. By contrast, the syntactic structure of the referring expressions used for these objects did show tailoring that suited their newness to the listener. In light of this evidence, it seems that tailoring may be more effective in higher-level structures, which are designed over longer intervals, than in individual word productions.

Given the above two issues, the generality of the listener-oriented account, especially regarding word-based intelligibility tailoring, is seriously limited. What's more, lexically conditioned variations add a new dimension to the computational complexity of monitoring and tailoring. In addition to other things, the listener model would also need to be informed

with relevant lexical information. Although such information can be stored in longterm memory and does not need online recalculation, it can only be retrieved in later stages of production, after lexical access, posing more challenge on the time efficiency of tailoring.

One may argue that listener orientation does not *require* accurate listener models or time-efficient tailoring, as speakers may tailor their production based on inaccurate estimation of listeners' needs and in ways that may not promote intelligibility. Nevertheless, this argument not only reduces the listener-orientation hypothesis to unfalsifiable conjecture, but also withdraws the motivation for proposing listener-orientation in the first place.

Thus, I will conclude that although listener orientation is a reasonable candidate for explaining pronunciation variation, the interpretation has some serious problems.

2.2.5.2 Speaker-oriented account

Words that are easier to produce are pronounced in a less articulated way while those that are harder to produce are pronounced in a hyperarticulated way. It seems obvious that there must be a causal relationship between ease (difficulty) of production and reduction (hyperarticulation). However, only half of the causal relation is well-grounded. As discussed in previous sections, disfluency signals trouble with planning and/or articulation, and speakers may slow down speech around disfluency in order to gain more time to solve the problems. Thus a natural causal relation can be established from difficulty in production to lengthening and hyperarticulation in general. What remains unexplained is the correspondence between ease of articulation and reduction. On a conceptual level, it seems equally reasonable to assume that words that are easy to pronounce will be reduced because speakers are more practiced or because they care less, or that easy words will be hyperarticulated because speakers have more resources (e.g. time) at hand for production.

Two accounts have attempted to connect production efficiency and pronunciation variation from both ends. Bybee (Bybee 2001, 2002a,b) attributed reduction in high-frequency forms to articulatory routinization: "If we take linguistic behavior to be highly practiced neuromotor activity [...], then we can view reductive sound change as the result of the automation of linguistic production" (Bybee 2002b:268). In other words, repeated production gives rise to more efficient neuromotor activity, which then leads to reduction and assimilation. This view can easily explain patterns observed in Phillips (1984)'s Frequency Actuation Hypothesis, since frequency-induced articulatory routinization will lead to physiologically-motivated sound changes and variations. It can also explain probabilistic effects for nonwords, since high-probability phonological sequences are mapped to articulatory gestures that are often practiced and their production will be easier, even if the sequence as a whole has never been produced before.

However, Bybee's articulatory routinization account fails to explain the full spectrum of the variation phenomena covered by Pattern B. A recent homophone study by Gahl (2008) has shown that even among words with the same phonological forms, pronunciation is still conditioned by word frequency. This evidence suggests that ease of articulation alone is not able to account for the full range of pronunciation variation. Another possible candidate,

which may also be a stronger candidate, is ease of access. As reviewed in previous sections, the major loci of most effects on production efficiency are in lexical access, which occurs before articulation (e.g. Griffin and Bock 1998; Jescheniak and Levelt 1994). When speakers are given enough time to retrieve the word form, some of the effects (e.g. frequency effects) tend to diminish (Balota et al. 1989; Jescheniak and Levelt 1994). These facts suggest that lexical access may be more important than articulation in facilitating speech production, and a separate connection needs to be built between lexical access and pronunciation variation, for a production-based theory to be convincing.

Bell et al. (2009) proposed an alternative hypothesis, specifically for linking lexical access and articulation. Bell et al. extended the finding of disfluency-related hyperarticulation to a more general mechanism which is also at work in fluent speech. This mechanism "helps coordinate lexical access and/or phonological encoding and the execution of the articulatory plan" (Bell et al. 2009:106), and it comes into play "when the specification of the current prosodic unit, most likely the phonological word (...), is slowed, but not so severely to require disfluent adaptations" (p106). By virtue of this coordination, words that are harder to access will be executed with longer durations. This account is very similar to the availability-based production theory (Ferreira and Dell 2000), which postulates that speech will slow down when the next unit is less available and causes problem in speech planning.

However, what is not clear is whether this synchrony mechanism can be generalized to explain variation among completely fluent words. Pierrehumbert (2002) suggests an across-the-board relationship between ease/difficulty of lexical access and degree of articulatory effort: "When a lexeme is retrieved and loaded into the phonological buffer, assume that a gradient value representing the ease of retrieval is passed to the buffer as a quantitative attribute of the Prosodic Word node. This parameter would control, or rather play a part in controlling, an overall parameter of articulatory clarity and effort." (Pierrehumbert 2002:107) Nevertheless, exactly how this parameter predicts for articulatory clarity and effort is not mentioned. Accumulated empirical evidence strongly implicates that words that are easy to access tend to undergo phonetic reduction, but the theoretical explanation has not been uncovered. Moreover, the opposite possibility has not been fully ruled out. As mentioned before, it is not unreasonable to assume that forms that are faster to access will be *hyperarticulated* because the speaker has more resources for producing them. A proposal of this kind has been made in Baese-Berk and Goldrick (2008; see discussion in §2.1.5), who attributed lengthening in VOT to higher activation levels (hence easier lexical access) during production. If this hypothesis is proved to be true, then one would expect to see an inverse U-shape relationship between ease of access and reduction, with words that are extremely hard or easy to access produced with less reduction (more hyperarticulation) and those in the middle with more reduction (less hyperarticulation).

In sum, although the accumulated empirical evidence strongly implies a causal relationship between ease/difficulty of production and degree of hypo/hyper-articulation, a crucial piece of evidence, regarding the relation between easy lexical access and speech reduction, is missing in the argumentation.

2.2.5.3 Confluence of listener-oriented and speaker-oriented accounts

As mentioned before, the two accounts are confluent for the factors in Table 2.3, because the factors that affect both perception and production have the same effects on both processes. As a result, the listener-oriented and speaker-oriented accounts generate identical predictions for patterns of pronunciation variation due to these factors. One thing following from the confluence is that we may be able to get away with less rigorous interpretation of the patterns, because if one theory doesn't seem to work, we can always resort to the other pattern. However, a careful examination of factors in Table 2.3 reveals that some of the factors that exhibit the confluence, such as frequency and repetition, are problematic for both accounts. For example, tailoring to perceptual difficulty in low-frequency words means that the speaker needs to track word-based perceptual differences, which adds to the computational burden of the listener model and challenges the time efficiency of tailoring. On the other hand, hyperarticulation in low-frequency words can also be attributed to difficulty of lexical access, but the production-based speaker-oriented account will have problem extending the relationship to words in high frequency ranges. As shown in this example, both the listener-oriented and speaker-oriented models will have problems accounting for the same variation phenomenon. Therefore, having a backup interpretation does not always save the theory, though it creates an illusion of doing so. Furthermore, the confluence may even harm the theories, since truly confluent patterns (which are always co-present and making the same predictions) cannot be tested independently. Judging from Table 2.3, this seems to be the case for lexical properties, if we assume that evidence for the facilitation of syntactic probability and orthographic regularity on production efficiency can be or has been found. In light of this, it will be desirable to find a lexical property for which listener- and speaker-oriented accounts can be separated. Ideally such a property should lead to opposite predictions under the two accounts, therefore the empirical findings can be unambiguously interpreted as evidence for only one of the speech models.

To sum up, this part of literature review on within-speaker pronunciation variation reveals both listener- and speaker-oriented forces in speech communication, but the magnitude and scope of these forces is not clear. Part of the problem is due to the seeming confluence of these forces, especially in lexically-conditioned pronunciation variation, which blurs the boundaries of these forces.

Before I end this section, it is necessary to mention that alternative theories of pronunciation variation also exist which do not immediately rely on listener- or speaker-oriented forces. These theories include the exemplar-based models of speech production (Pierrehumbert 2002; see discussion before in §2.1.5) and various information-theoretic theories (the Probabilistic Reduction Hypothesis, Jurafsky et al. 2001b; the speech efficiency hypothesis, Van Son and Pols 2003; the Smooth Signal Redundancy Hypothesis, Aylett and Turk 2004; the Universal Information Density theory, Frank and Jaeger 2008; Levy and Jaeger 2007; see discussion before in §2.2.3).

An exemplar-based model of speech production accounts for variation in pronunciation via mechanisms of offline restructuring of the lexicon by perception. If some words are hard

to perceive, hyperarticulated tokens will have a greater chance of being correctly recognized and therefore remembered as exemplars. Thus, the cloud of exemplars for perceptually challenging words will be skewed toward a high-density cluster of hyperarticulated tokens. Accordingly, during production, when the speaker draws from existing exemplars, hyperarticulated tokens are more likely to be selected and produced. Because the exemplar-based theory relies on delayed but permanent responses in production to perceptual difficulty, it relaxes the constraint on online monitoring and speech tailoring. However, this theory comes to problems when accounting for effects that depend on the context, such as predictability and syntactic probability, which cannot be stored in longterm lexical representations.

In information-theoretic models of speech production, pronunciation variation is attributed to the differences in information load and/or entropy. Utterances are designed in a way that lexical units that contain more information (or less expected information) will be hyperarticulated and those that contain less information (or more expected information) will be reduced. These hypotheses resemble those in listener-oriented models, though information-theoretic models do not make explicit references to deliberate monitoring for and tailoring to listener's needs. Information-theoretic models can readily explain various probabilistic effects (which can all be related to predictability) on pronunciation variation in normal context. However, it is not clear how current versions of information-theoretic models would predict for situations where expectedness to the speaker differs from that to the listener. Moreover, variations conditioned by *form* instead of by *meaning*, as in the case of neighborhood-conditioned variation, are largely unaccounted for in information-theoretic models.

In sum, these alternative models of speech production are not superior to the listener-oriented model and the speaker-oriented model that were presented before. Moreover, the source of variation in these alternative models is (implicitly) attributed to perceptual difficulty and information predictability, both of which are contained in the listener-oriented and speaker-oriented models, though the specific mechanisms leading to pronunciation variation are different. Therefore, I will consider the listener-oriented model and the speaker-oriented model as two basic models of speech production for future discussion in this thesis. References to alternative models will only be made when necessary.

In the next section, I will formulate two sets of hypotheses, from the two basic models, regarding the effects of neighborhood structure on pronunciation variation.

2.3 Current work

2.3.1 Goals

The ultimate goal of the current study is to understand the effects of neighborhood structure on pronunciation variation. In doing so, it is also my hope to provide insights into the listener-oriented and speaker-oriented natures of speech production. As reviewed in the previous section, the literature on pronunciation variation strongly suggests that

both ease/difficulty of perception and ease/difficulty of production can cause pronunciation variation, and the underlying mechanisms can be sought from listener-oriented and speaker-oriented models of speech production. However, neither model is able to account for all empirical evidence, and the problems are exacerbated by the confluence of the two accounts in most variation phenomena, especially in lexically-conditioned variations.

Thus, the broader literature review on pronunciation variation points back to phonological neighborhoods, because (a) neighborhood characteristics are lexical properties and (b) neighborhood characteristics have opposite effects on perception and production efficiency. The second property of neighborhood characteristics is both unique and crucial. In fact, neighborhood density (and neighbor frequency) may be the only known factor that produces opposite effects on perception (*inhibition*) and production (*facilitation*). Because of this property, neighborhood density provides a crucial testing ground for separating listener-oriented and speaker-internal forces in speech production. If high-density words indeed exhibit hyperarticulation, it will provide evidence for the connection from perception to pronunciation variation, which will in turn support the listener-oriented model. On the other hand, if high-density words exhibit speech reduction, it will provide evidence for the connection from production efficiency to pronunciation variation, supporting the speaker-internal model. In both cases, the interpretation will be unambiguous, because at most one of the two possible accounts will be consistent with the findings. Thus, the investigation of neighborhood effects on pronunciation variation will improve our knowledge about models of speech production in general.

A secondary goal of the current research is to examine neighborhood-conditioned variation *in context*. Compared with previous literature on neighborhood effects in production, a distinguishing feature of the current research is the use of conversational speech data. As the field stands right now, most of the evidence comes from word reading experiments, with the exception of Scarborough (2004, 2010), who inspected neighborhood effects in connected speech elicited in laboratory experiments. To my best knowledge, no study has investigated neighborhood effects in spontaneously produced conversational speech. It is highly likely that neighborhood effects in spontaneous speech will be very different from the previously observed neighborhood effects in read speech. For one thing, the constant presence of a listener in natural conversation may help motivate listener adaptation in speech production. On the other hand, the complicated task of speech planning in unscripted speech may induce stronger speaker-oriented effects. Thus, the investigation of neighborhood effects in spontaneous speech production may yield highly interesting results.

In addition, the current research is also intended to further our knowledge about the relationship between duration and vowel dispersion. As mentioned in previous sections, these two phonetic features are most well-studied in pronunciation variation, and are generally assumed to go hand in hand. However, cases of dissociation have also been documented in the literature. Specifically, if neighborhood-conditioned speech variation is indeed driven by both listener-oriented and speaker-internal forces, toward opposite directions, it is not unlikely to see dissociation among different aspects of speech production.

2.3.2 Hypotheses

In the current research, two corpus studies will be conducted to examine neighborhood effects in conversational speech production. The first study concerns variation in word duration, and the second study concerns variation in vowel dispersion. Following from previous literature on phonological neighborhoods and pronunciation variation, two sets of hypotheses are formulated (see below). The first set of hypothesis ("speaker-oriented hypothesis") attributes variation to speaker-internal properties, while the second hypothesis ("listener-oriented hypothesis") considers variation as a result of listener orientation. Neighbor frequency is assumed to have exactly the same effects as neighborhood density, though there is more empirical evidence for neighborhood density in the current literature. Besides, as discussed in previous sections, individual arguments (e.g. 1b) in the hypotheses may lack theoretical support. Therefore, the hypotheses represent the *most likely* scenarios given what is known in the current literature, and should not be taken as rigorous predictions.

(1) Speaker-oriented hypothesis

   (a) High-density words are easier to produce than low-density words. Words with high-frequency neighbors are easier to produce than those with low-frequency neighbors.

   (b) Easy-to-produce words tend to have shorter durations and more reduced vowels.

   (c) Therefore high-density words and words with high-frequency neighbors will be realized with *shorter* durations and *more reduced* vowels than low-density words and words with low-frequency neighbors, respectively.

(2) Listener-oriented hypothesis

   (a) High-density words are harder to recognize than low-density words. Words with high-frequency neighbors are harder to recognize than those with low-frequency neighbors.

   (b) Words with longer durations and more dispersed vowels are more intelligible.

   (c) Speakers monitor for listeners' needs and modify their speech accordingly.

   (d) Therefore high-density words and words with high-frequency neighbors will be realized with *longer* durations and *more dispersed* vowels than low-density words and words with low-frequency neighbors, respectively.

Chapter 3

Methods

This dissertation contains two corpus studies. The first one explores the effect of phonological neighborhoods on word duration and the second one explores the same effect on vowel dispersion. The methods in the two studies are highly similar. Both studies were conducted on spoken tokens of CVC monomorphemic content words extracted from an English speech corpus. Mixed-effect linear regression models were used to investigate the critical effects while controlling for other factors. In view of the similarity of the two corpus studies, this chapter describes methods that are common to the two studies. Detail in data and models that is specific to each study will be presented in the relevant sections of Chapters 4 and 5.

3.1 Data

3.1.1 Corpus

All speech data used in the current work came from the Buckeye Speech Corpus (Buckeye corpus; Pitt et al. 2007), which contains about 300,000 words of interview speech from 40 speakers recruited from local communities in Columbus, OH. All speakers were middle-class Caucasians who were either born in or around Columbus or moved there no later than age 10. Age (below or over forty) and sex were balanced when recruiting speakers. Each speaker participated in one free-form interview session that lasted between 30 to 60 minutes, with one experimenter. All interviews were completed by Spring, 2000. For more information about data collection in the corpus, please refer to the corpus manual (Kiesling, Dilley, and Raymond 2006).

Two types of transcription were provided in the corpus, mainly for the speaker's speech[1]: (a) an orthographic transcription for all words and word fragments (e.g. *th- they*); (b) time-aligned phonetic transcription at both the word level and the phone level. The second type of transcription is a major source of data for the current work.

[1]The experimenter's speech, though sometimes audible from the recording, was not transcribed in words.

3.1.2 Database

For the purpose of the current work, I extracted all consonant-vowel-consonant (CVC) monomorphemic content words. The selection of target words (tokens) is explained in more detail below.

3.1.2.1 Type-based exclusion

To control for word length, stressability, morphological and CV structure, I restricted the analysis to monomorphemic CVC content words. The complete word list was compiled using the following procedure.

First of all, an automatic script was run through the corpus transcription to find all instances of words that are associated with CVC pronunciations in the citation forms (altogether 839 words represented by 74,073 tokens). It should be noted that since the current version of the corpus transcription has no part of speech tagging or syntactic information, the concept of "word" in this work is purely based on orthography. So *come* and *came* are recognized as different words, but *can* (aux.) and *can* (n.) are merged, and so are *bush* and *Bush*, since all words were transcribed with lower case letters.

Secondly, morphologically complex words were removed, including 24 contracted forms (e.g. *he'll*) and 46 morphologically complex words (e.g. *goes*), leaving only base forms (e.g. *bag*) and irregular inflected forms (e.g. *came*) on the list.

Thirdly, 74 function words were removed, to control for stressability. The identification of function words was based on both the function/content distinction in the Hoosier mental lexicon (HML; Nusbaum et al. 1984) and the part of speech information from the CELEX lexical database (Baayen et al. 1993). To ensure accuracy, the resulting word list was hand-checked. A word was identified as a function word if it is (a) listed as a function word in HML, **or** (b) listed in CELEX with a primary syntactic category that does not belong to the four categories for content words (adjective, adverb, noun, verb). Besides, word forms corresponding to multiple lemmas (e.g. *can*) were considered to be function words as long as one of the lemmas is a function word. Altogether there were three words (*can, might* and *will*) of this kind in the Buckeye corpus. One may suggest that the content use of these words should be retained in the database. The major reason why this approach was not adopted is because the corpus does not come with syntactic transcription and therefore an automatic program for semantic disambiguation is difficult to implement. Considering that these forms are much more often used as function words than as content words, removing all tokens of these words seems to be an acceptable strategy under the practical constraints. Compared with words that remained in the final database (range of ND is [3,40]), these three words are relatively high in neighborhood density (ranging from 28 to 35, according to the Hoosier mental lexicon). But due to the small number of these words (n=3), the removal should not cause a significant change in the distribution of ND in the final word list (n>500). Finally, the word *cause* was also removed from the word list, despite being listed as a content word in both dictionaries, because the Buckeye transcription used the same orthographic form for

the content word *cause* and the reduced form of the function word *because*.

In the next step, 9 words were removed because they are frequently used as discourse fillers in conversations. Examples are *like* and *mean* (as in *I mean*). Similarly, 8 words that are often used as interjections (e.g. *man* and *dude*) were also removed. The removal of these words is also based on consideration of stress, since discourse fillers and interjections tend to have extraordinary prosodic patterns compared to other words. Neighborhood density of these words ranges from 9 to 35, which is about the same range as words in the final database.

Five more words (e.g. *tear* and *route*) were excluded for having multiple phonemic representations in standard American English, because there is no systematic way to find out which pronunciation was intended by the speaker , which makes it hard to determine the neighborhood of the phonetic target.

The next phase of type-based exclusion concerned the availability of lexical measures in the dictionaries. As §3.2.1 will discuss, a wide range of independent variables were included in the model, and the coding of these variables also led to data exclusion. In particular, 46 words were excluded for missing part of speech in the CELEX database. Most of these words are proper names (e.g. *Pitt*) and technical jargon (e.g. *DOS*). Thus, in addition to part of speech, they also tend to miss other lexical properties such as frequency and familiarity. It should be noted, however, that the current database does include words that are ambiguous between proper names and common nouns (e.g. both *Bush* and *bush* are transcribed with *bush*), and no effort has been made to disambiguate such cases.

In addition, 31 words were removed due to missing neighborhood metrics from the Hoosier mental lexicon and 2 words were removed for missing frequency from the CELEX database.

After all the above steps of type-based exclusion, the database was left with 594 words represented by 16,660 tokens. Table 3.1 summarizes the number of words (tokens) excluded at each step.

3.1.2.2 Token-based exclusion

Further data exclusion was carried out on specific tokens of included word types. Token-based exclusion was mostly motivated by contextual properties. Specifically, 3,058 tokens were removed for being at stretch-initial or stretch-final positions. Not only do these tokens lack contextual variables such as speech rate and bigram probability, they are also prone to phrase-boundary intonations. Furthermore, 360 tokens with bigram probability equal to 1 (i.e. completely predictable given the preceding or following word) were also removed, in order to avoid fixed expressions and hapax legomena.

The final database consists of 13,242 tokens of 540 unique words from 40 speakers[2].The complete word list, as well as the list of excluded words, is given in Appendix A. The model for word duration used the full database while the model for vowel dispersion made use of

[2]Token-based exclusions have accidentally caused 594-540 = 54 type exclusions, mostly due to the exclusion of tokens of low-frequency words.

Reason for exclusion	**Number of word types removed**	**Number of word tokens removed**
Contracted forms	24	2675
Morphologically complex words	46	809
Function words	74	44480
Discourse fillers	9	8494
Interjections	8	260
Multiple pronunciation	5	292
Missing part of speech	46	96
Missing neighborhood density	31	291
Missing frequency	2	16
Remainder	*594*	*16,660*

Table 3.1: Number of words (tokens) removed at each step of type-based exclusion.

a subset of the database as appropriate for formant analysis. Specific detail about the vowel data set will be presented in the data section of Chapter 5.

3.2 Mixed-effect models

The statistical tool used in both corpus studies is mixed-effect linear regression models, which is an extension of the widely used linear regression analysis. In the following, I will briefly introduce the design and construction of a mixed-effect model, as well as methods for interpreting model results and evaluating model reliability.

3.2.1 Model design

A linear regression model is built to find linear relationships between an outcome variable and some explanatory variables. For example, if we are interested in the effect of aging on speech rate, we can measure the speech rates of 100 people of different ages and fit a linear equation to the observed data. The equation can be expressed as $\texttt{rate} = \alpha + \beta \cdot \texttt{age}$. If the coefficient ($\beta$) is found to be positive, it means that speech rate is higher among older people, whereas a negative coefficient would indicate the opposite. If β is not significantly different from zero, that would suggest that age does not affect speech rate.

However, in the example above, the model fails to capture a crucial type of variance, that is, individual differences among speakers. It is possible that people in general tend to speak slower as they get older, but their initial speech rate can be different. Thus if we

compare the speech rate of a 70-year-old person with that of a 50-year-old person, the older person may actually speak faster than the younger one, though they both talk slower than when they were younger. Such individual variance must be treated as random in nature and cannot be modeled as a regular predictor like `age`, but without knowledge of it, the model will be trying in vain to find a relationship between speech rate and absolute age.

To overcome this problem, we will first need to collect more data points from each speaker, at different ages. Second, we will also need a more sophisticated model, one that can handle both regular trends and random variance. This is where mixed-effect models (Hartley and Rao 1967) become useful. A mixed-effect model can accommodate two types of predictors: fixed-effects predictors and random-effects predictors. A fixed-effects predictor is fitted with a coefficient, whereas a random-effects predictor is assigned with normally distributed adjustments (with a mean of zero) for all individuals in the sample. Using our previous example, a mixed-effect model on speech rate will contain `age` as a fixed-effects predictor and `speaker` as a random-effects predictor. The predicted value for the speech rate of speaker i can be expressed as the following: $\texttt{rate} = \alpha + \beta \cdot \texttt{age} + \eta_i$, where η_i is the speaker-specific adjustment.

In the current work, two mixed-effect models were designed with similar structure, one with word duration as the outcome variable and the other with degree of vowel dispersion as the outcome variable. In both models, neighborhood density and neighbor frequency were entered as fixed-effects predictors, as well as a wide range of control factors, regarding the word, the context and the speaker. In order to control for speaker- and word-specific differences in pronunciation, `speaker` and `word` were entered as crossed random-effects terms regarding the intercept. The general formula of the models can be expressed as the following:

$$Y = \alpha + \sum \beta_k X_k + \eta + \gamma, \tag{3.1}$$

where Y is the outcome variable, α is the intercept, β_k is the coefficient of the kth fixed-effect term X_k, $\sum \beta_k X_k$ is the sum of all fixed-effects terms, and η and γ are by-speaker and by-word adjustments, respectively.

It might seem redundant to include lexical and speaker characteristics as fixed-effects terms if the model already provides individual adjustments for different words and speakers. However, the point of doing so is to capture systematic variation in the data so that regular trends will not be attributed to random variance. In fact, even in cases where adding fixed-effects predictors does not result in a better model fit, it may still be valuable because the size of random adjustments may be reduced.

3.2.2 Model construction

All mixed-effect models in the current work were constructed using the `lmer` function in the **lme4** package (Bates and Maechler 2010) in R (version 2.11.1). Following Baayen (2008), all numerical predictors were centered (i.e. with the mean value removed) in order to reduce collinearity in the model.

The modeling strategy used in the current work was to construct a series of models before arriving at the final model. The purpose of doing so is to filter out insignificant predictors and possible outliers in the observed data[3]. The procedure for model construction is described in the following.

First, a full model is created with all random- and fixed-effects terms. Second, judging from the summary of model parameters, clearly insignificant fixed-effects predictors are eliminated. A subsequent model is built with the remaining predictors. In the next round of elimination, predictors with borderline significance are selected for further tests by model comparison. If removing a certain predictor does not cause a significant change in model fit, the predictor will be considered as insignificant and removed from the next stage of model construction. After all suspected predictors are independently tested, a new model is built without the insignificant predictors, and thus completes the predictor-trimming process. A small set of predictors are exempted from this process, including the crucial factors (neighborhood density and neighbor frequency) and their close correlates (phonotactic probability).

The resulting model is further examined for the existence of outliers in the data. Possible outliers are identified by inspecting the distribution of model residual. One of the assumptions underlying the models used here is that error is normally distributed. Departures from normality of model residuals would indicate possible violations of that assumption and warrant further investigation. I will turn to this point in the discussion of each model.

3.2.3 Model interpretation

Model interpretation includes three parts: general model fit, random effects and fixed effects. Mixed effect models are fitted with a technique called RELATIVIZED MAXIMUM LIKELIHOOD. The major indicator of the goodness of model fit is log likelihood. An alternative measure is the proportion of explained variation (R^2) in the data, which is calculated as the square of the correlation between the observed outcomes and the predicted values.

For random effects, the most informative parameter is the variance of individual adjustments, which indicates how big the average adjustment is (because they are centered around zero). The interpretation of fixed-effects terms is more complicated. The most important statistics for fixed-effect terms are the estimated coefficients (β) and their associated standard errors and t values. The smaller the standard error is (relative to the coefficient) and the greater the absolute t value is, the more significant the effect is. However, the current version of the **lme4** package does not produce p values for t and F tests, because it is still unclear how to appropriately calculate degrees of freedom for mixed effect models.

One simplified solution is to use the upper bound of degrees of freedom (i.e. the number of observations subtracted by the number of fixed-effects predictors), and this works reasonably well for large data sets with thousands of observations. In the current work, a model typically has near or over 10,000 data points and a dozen predictors. Using the upper bound

[3]Though the inclusion of all control factors in the model is theoretically motivated, it is not necessarily the case that they will all turn out to be significant in the current model, and having insignificant predictors increases the amount of noise in model prediction.

of degrees of freedom ($\approx$ 10,000), an absolute t value of 2.5 will correspond to a p value of 0.01, whereas $|t|$=2 corresponds to p=0.05, and $|t|$=1 corresponds to p=0.31. Thus I consider an absolute t value smaller than 1 as an indicator of lack of significance in the current work, and an absolute t value between 1 and 2 as an indicator of borderline significance. For predictors with absolute t values grater than 2, I further examine the significance of these predictors using model evaluation techniques (see the section below).

If a predictor is significant, then the sign of its coefficient indicates the direction of the effect. A positive coefficient suggests that the outcome variable will increase when the predictor increases (or changes from the reference level, for categorical predictors), while a negative coefficient suggests the opposite.

The size of an effect is related to both its coefficient and the range of variation in the predictor variable. In practice, it can be gauged by calculating the predicted range of variation in the outcome variable by just varying the critical predictor and keeping all other predictors at their mean levels.

3.2.4 Model evaluation

Given the complexity of the model structure and the large size of the data set in the current work, it is important to examine the reliability and robustness of model results. A number of model evaluation techniques were used in the current work, including model comparison, MCMC-based parameter evaluation, partial effects inspection and cross validation. A brief description of the evaluation techniques is given below.

3.2.4.1 Model comparison

An effective way of examining the significance of a certain predictor is to compare the fit of the model with and without the critical variable. If removing a predictor causes no significant change in model fit, then the predictor is probably not making important contribution to the prediction for the outcome variable.

3.2.4.2 MCMC-based parameter evaluation

Apart from using the upper bound of degrees of freedom, an additioal way of interpreting t values is to use the Markov chain Monte Carlo (MCMC) sampling technique. This technique works by sampling a large number of estimates for model coefficients, which gives insight into the posterior distributions of the parameters. P values can also be generated based on the MCMC samples. In the current work, I used the `pvals.fnc` function in the **lme4** package for MCMC-based analysis, which by default generates 10,000 samples and reports the lower and upper bounds for the 95% highest posterior density (HPD) intervals.

3.2.4.3 Cross validation

Cross validation is used to evaluate the generalizability of model findings. A cross validation procedure simulates the effect of having different data sets by building models on randomly selected subsets of the data. If all the models generated from partial data sets find similar results and converge around the model based on the complete data set, one will be able to say that the observed effects persist across subsets of the data, which increases our confidence in the generalizability of the current findings to unseen data.

3.3 Chapter summary

In this chapter, I have introduced the methods that were used in the current work. Both corpus studies, one on word duration, and the other on degree of vowel dispersion, used a word/token database compiled from the Buckeye corpus. Mixed-effect linear regression models were constructed for studying the effects of neighborhood density and neighbor frequency on the variations of word duration and degree of vowel dispersion (modeled separately). In the following two chapters, I will present the two corpus studies in turn.

Chapter 4

Phonological Neighborhood Density and Word Duration

This chapter presents the first corpus study in the thesis. The goal of the study is to investigate the effect of phonological neighborhood structure on word duration, comparing a speaker-oriented hypothesis and a listener-oriented hypothesis. The speaker-oriented hypothesis predicts that words from dense neighborhoods will be realized with shorter durations than words from sparse neighborhoods, while the listener-oriented hypothesis predicts the opposite to be true.

To distinguish the two hypotheses, a mixed-effect linear regression model is constructed on word token duration. Input data are spoken tokens of CVC monomorphemic content words from the Buckeye corpus. Neighborhood density and neighbor frequency, as well as other factors that are known to affect word duration in conversational speech, are entered into the model as predictors. To preview the results, the model reveals a shortening effect of neighborhood density over and above other known predictors of word duration, indicating that high-density words are shorter than low-density words, as predicted by the speaker-oriented hypothesis.

The organization of the current chapter is as follows: Section 4.1 recapitulates the two hypotheses; Section 4.2 describes the database and the variables in the model; Section 4.3 presents modeling procedure and model results; Section 4.4 discusses the results from alternative models; Section 4.5 ends the chapter with a brief discussion of current findings.

4.1 Predictions

As discussed in Chapter 2, neighborhood density has been found to have opposite effects on perception and production. In the realm of perception, words from dense neighborhoods take more time to recognize and the recognition accuracy is lower (e.g. Luce and Pisoni 1998), whereas in production, high-density words can be initiated more quickly production and the naming accuracy is higher (e.g. Vitevitch 2002).

Given the facilitatory and inhibitory effects of neighborhood density, two hypotheses can be formed regarding regarding how neighborhood density might affect pronunciation variation. A speaker-oriented hypothesis predicts that high-density words will be more reduced than low-density words, as a natural result of facilitated production. Contrarily, a listener-oriented hypothesis predicts that high-density words should be hyperarticulated, in order to compensate for the perceptual difficulty. Thus phonological neighborhood density presents a rare testing ground for separating the contribution of speaker- and listener-oriented forces in speech communication. In terms of duration, the speaker-oriented hypothesis predicts shorter durations in high-density words whereas the listener-oriented hypothesis predicts the opposite to be true.

The complete arguments of both hypotheses are shown in (1) and (2) (repeated from Chapter 2). For the time being, it is assumed that neighbor frequency operates in the same way as neighborhood density.

(1) Speaker-oriented hypothesis regarding neighborhood effects on wordduration

   (a) High-density words are easier to produce than low-density words. Words with high-frequency neighbors are easier to produce than those with low-frequency neighbors.
   (b) Easy-to-produce words tend to be realized with shorter durations.
   (c) Therefore high-density words and words with high-frequency neighbors will be *shorter* in duration than low-density words and words with low-frequency neighbors, resepectively.

(2) Listener-oriented hypothesis regarding neighborhood effects on wordduration

   (a) High-density words are harder to recognize than low-density words. Words with high-frequency neighbors are harder to recognize than those with low-frequency neighbors.
   (b) Words with longer durations are more intelligible.
   (c) Speakers monitor for listeners' needs and modify their speech accordingly.
   (d) Therefore high-density words and words with high-frequency neighbors will be *longer* in duration than low-density words and words with low-frequency neighbors, respectively.

4.2 Data

4.2.1 Database

The database for the current study contains 13,242 spoken tokens of CVC monomorphemic content words from the Buckeye corpus, featuring 40 speakers and 540 word types. Detailed description of the database was given in Chapter 3. On average each speaker has provided 331.05 tokens (s.d.=131.30) and 107.92 word types (s.d.=22.53).

4.2.2 Coding variables

The outcome variable of the model is log word token duration (log(s)). Other variables include the two crucial variables (neighborhood density and neighbor frequency) and a large set of lexical, contextual and sociolinguistic control factors (baseline duration, bigram probability, disfluency, familiarity, frequency, orthographic length, part of speech, phonotactic probability, previous mention, speaker age, speaker sex and speech rate). The inclusion of most control variables is motivated by previous research on pronunciation variation, as reviewed in Chapter 2. Below is a list of descriptions of all the variables that are included in the model. Continuous predictor variables are all log transformed before entering the model.

Word token duration is calculated based on the beginning and end time of the token, as transcribed in the corpus. Log transformed word duration is used in the model.

Neighborhood density and neighbor frequency are the critical predictor variables in the current model. Neighborhood density is the number of phonological neighbors under the one-phoneme difference rule. Neighbor frequency is the mean log frequency of neighbors. Both measures come from the Hoosier mental lexicon (Nusbaum et al. 1984), which is in turn based on pronunciations from the Webster pocket dictionary (which has about 20,000 lexical entries) and word frequency from Kučera and Francis (1967) (which is based on the 1-million word Brown corpus).

Baseline duration is an estimate for word duration based on the phonemic segments of the word. Some phonemes are inherently longer than others (e.g. [iy] is longer than [ih][1]), and such phoneme-specific length differences should be considered by the model. Raw baseline duration is calculated as the sum of the mean lengths of the phonemes in the word's citation form. For instance, raw baseline duration of the word *cat* is the sum of the mean length of [k] plus the mean length of [ae] and the mean length of [t]. Mean segmental lengths are averaged over the entire Buckeye corpus. Raw baseline duration is log transformed before entering the model.

Bigram probability is calculated as the conditional probability of the current word given the adjacent word in the Buckeye corpus, as an estimate for contextual predictability. Each word token is coded with two bigram probability measures: one with the preceding word (i.e. $C(W_{n-1}W_n)/C(W_{n-1})$) and the other with the following word (i.e. $C(W_nW_{n+1})/C(W_{n+1})$). Frequency counts of the neighboring words and the word pairs are based on the entire corpus. Both probability measures are log transformed.

Importantly, only adjacent words in the same stretch are considered as valid context. Thus words at the beginning or end of a stretch will have no bigram probability. The end of a stretch is encountered when one of the following items appears in the transcription: silence longer than half a second, cutoff words (i.e. transcribed with <CUTOFF>

[1]To be consistent with the Buckeye transcription, phonetic symbols in this dissertation are all transcribed in ARPABET.

labels), words with lexical or phonological errors (<ERROR>), interviewer's speech (<IVER>), and non-linguistic sounds such as laughter (<LAUGH>), non-speech vocal noise (<VOCNOISE>) and environmental noise (<NOISE>).

Disfluency is coded by two logical variables, one for the preceding context and the other for the following context. If the target token is immediately preceded or followed by cutoff words (i.e. transcribed with the <CUTOFF> labels), word errors (i.e. transcribed with the <ERROR> labels) or filled pauses (*um, uh* or *you know*), the corresponding disfluency variable will be coded as TRUE, otherwise it will be FALSE.

Familiarity is the subjective familiarity rating of words. The current metric is from the Hoosier mental lexicon, which encodes word familiarity from 1 (unknown) to 7 (highly familiar).

Frequency is the usage frequency of the word form. The current measure is from the CELEX frequency dictionary for wordforms, which is in turn based on the 17.9 million-word Collins Birmingham University International Language Database (COBUILD corpus; Sinclair 1987). Log frequency (i.e. $ln(Freq + 1)$) is used in the model.

Orthographic length is the length of the word in letters.

Part of speech is the syntactic category of the target word. Since the Buckeye corpus has no syntactic transcription, I used part of speech tags from the lemma-based CELEX syntax dictionary and all tokens of the same word are coded with the same part of speech. For category-ambiguous words, usage frequency for each possible part of speech is summed and the most frequent part of speech is recorded for the word (e.g. *time* is coded as a noun instead of a verb). Part of speech of irregular forms were manually added (e.g. *came* is coded as a verb; all past participles are coded as verbs). Words in the final database all belong to one of the four syntactic categories: adjective (A), adverb (ADV), noun (N) and verb (V).

Phonotactic probability refers to the probability of having the exact phonemic composition of a word. In this work, each word is coded with two probability measures: average phoneme probability and average biphone probability. Take the word *cat* for example. The average phoneme probability of *cat* is the mean value of the probability of having [k] in the initial position, the probability of having [ae] in the second position and the probability of having [t] in the third position in the English language. Likewise, average biphone probability of *cat* equals the mean value of the probability of having [k ae] in the first and second positions and the probability of having [ae t] in the second and third positions. All raw probability measures are obtained from the web-based Phonotactic Probability Calculator (Vitevitch and Luce 2004). Log probabilities are used in the model.

Previous mention is coded as a logical variable. If the same word has been produced by the speaker at least once from the beginning of the interview to the point right before the target token, it will be coded as TRUE, otherwise it will be FALSE.

Speaker age is coded as a binary variable in accordance with the age stratification in the corpus. Exactly half the speakers (n=20) were under forty years old (Y) and the other half were above (O) at the time of recording. The actual age of the speakers ranged from late teens to late seventies.

Speaker sex is coded as a binary variable. Half the speakers are male (M) and the other half are female (F).

Speech rate is coded by two numerical variables, one for the preceding part of the local stretch and the other for the following part. To avoid autocorrelation, the target token is *not* included in the calculation of either speech rate measure. Raw speech rate is measured in number of syllables per second. Log transformed speech rate is used in the model.

Table 4.1 shows the summary statistics for all categorical variables and Table 4.2 summarizes for numerical variables.

Variable	**Token counts at each level**
`disfluency (before)`	F = 13096; T = 146
`disfluency (after)`	F = 12555; T = 687

`part of speech`	adjective (A) = 2582;
	adverb (ADV) = 528;
	noun (N) = 4953;
	verb (V) = 5179

`previous mention`	F = 3840; T = 9402
`speaker age`	O = 7492; Y = 5750
`speaker sex`	F = 6299; M = 6943

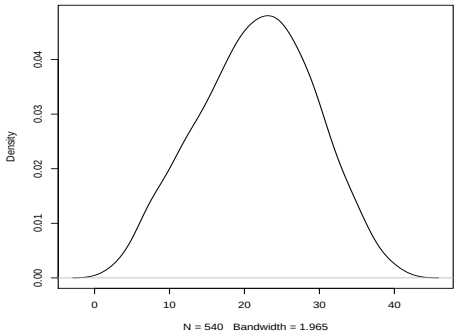
Table 4.1: Summary statistics for categorical variables based on the the 13,242 tokens in the database.

Variable	Min	Max	Median	Mean	s.d.	Log
word token duration	0.010 (s)	1.04 (s)	0.25	0.26	0.097	Yes
neighborhood density	3	40	21	20.61	6.84	
neighbor frequency [2]	1.26	3.17	2.07	2.07	0.25	
baseline duration	0.19 (s)	0.38 (s)	0.25	0.25	0.034	Yes
bigram probability (before)	0.000079	0.75	0.0050	0.027	0.070	Yes
bigram probability (after)	0.000079	0.83	0.0041	0.029	0.077	Yes
familiarity	2.4167	7	7	6.95	0.14	
frequency	0 (per 18m)	49655 (per 18m)	6687	9871.19	11079.75	Yes
orthographic length	3 (letter)	7 (letter)	4	4.05	0.71	
phonotactic probability (phoneme)	0.012	0.098	0.046	0.048	0.016	Yes
phonotactic probability (biphone)	0.0002	0.016	0.0022	0.003	0.0023	Yes
speech rate (before)	0.95 (syll/s)	33.33 (syll/s)	5.91	6.21	2.26	Yes
speech rate (after)	0.42 (syll/s)	41.04 (syll/s)	5.19	5.28	1.71	Yes

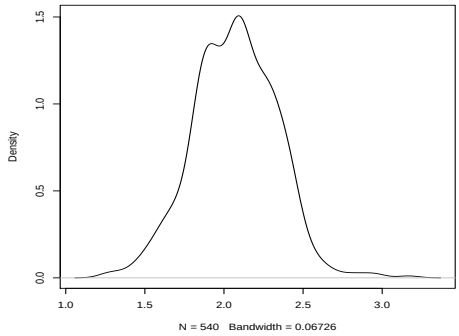
Table 4.2: Summary statistics for numerical variables (in raw values) based on the the 13,242 tokens in the database.. "Log" denotes whether the variable is log transformed before entering the model.

The next set of figures (4.1 and 4.2) shows the distribution of numerical variables in the database. Because some variables are properties of the word (e.g. frequency) while others are properties of the token (e.g. word token duration), a distinction is made between type-based and token-based variables. For type-based variables, distribution over the word set is shown, while for token-based variables, distribution over the token set is shown. If a variable is log transformed, distributions of both the raw values and the log values are shown. As can be seen in the figures, most durational and probabilistic variables achieve more normal distributions after log transformation.

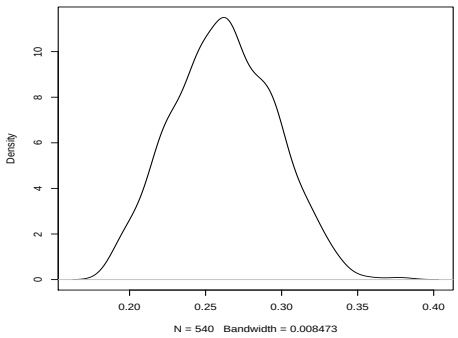
[2] Average neighbor frequency is already log-transformed in HML.



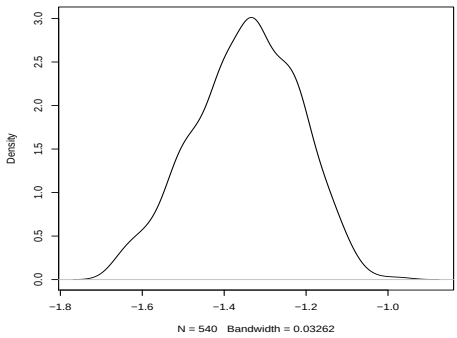
(a) Neighborhood density



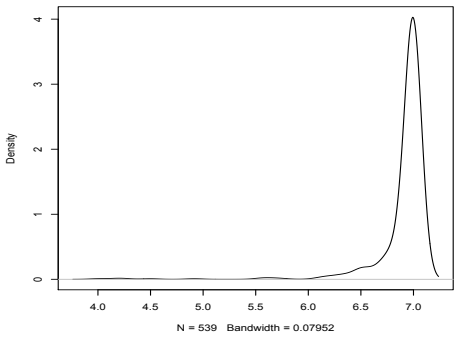
(b) Neighbor frequency



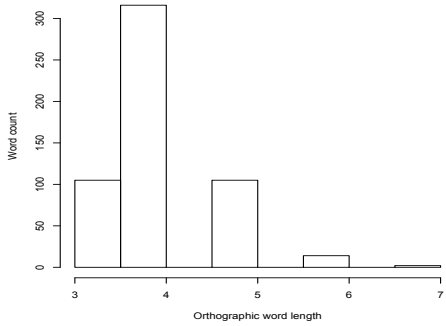
(c) Raw baseline duration



(d) Log baseline duration

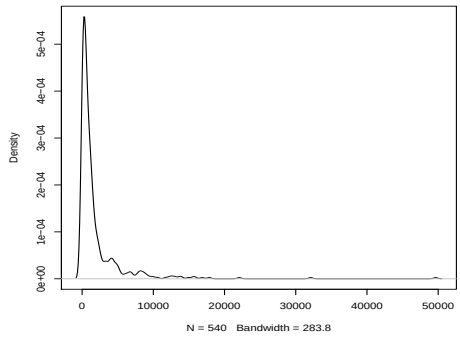


(e) Familiarity. (For clearer presentation of the distributive pattern, an outlier datapoint, with a familiarity rating below 3, was not included in the density plot.)

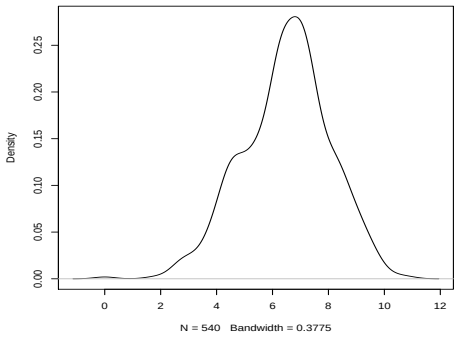


(f) Orthographic length

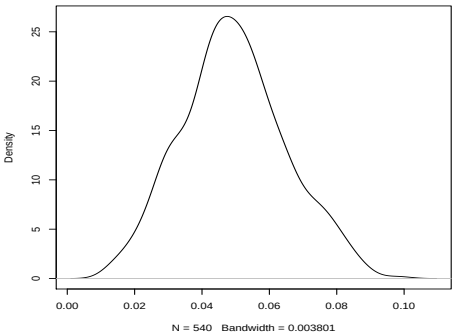
Figure 4.1: Distribution of type-based variables in the word duration model. Distributions are based on word types (n=540). For orthographic length, histogram is shown. For all other variables, probability functions are shown. (Continued on the next page.)



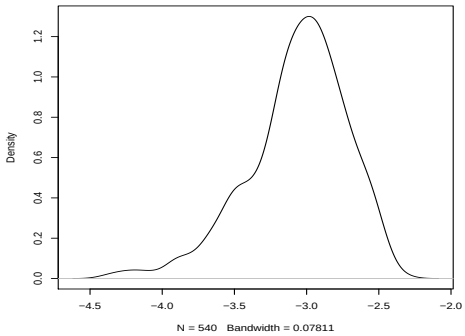
(g) Raw frequency



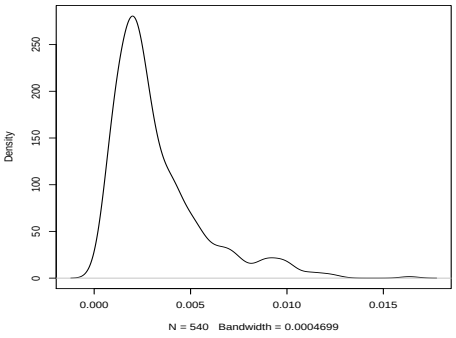
(h) Log frequency



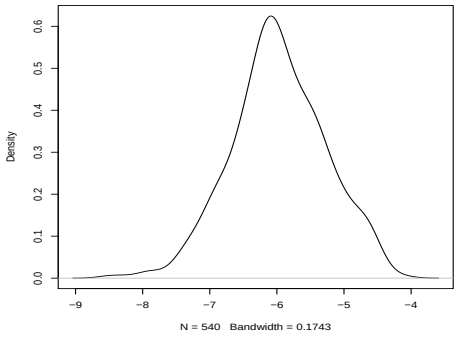
(i) Raw phonotactic probability (phoneme)



(j) Log phonotactic probability (phoneme)

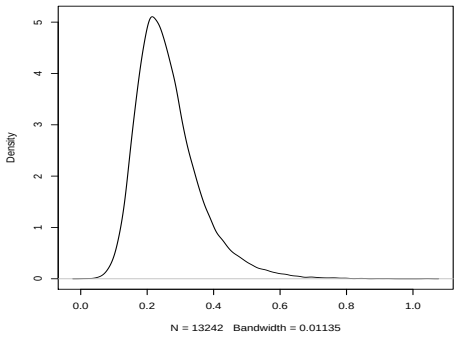


(k) Raw phonotactic probability (biphone)

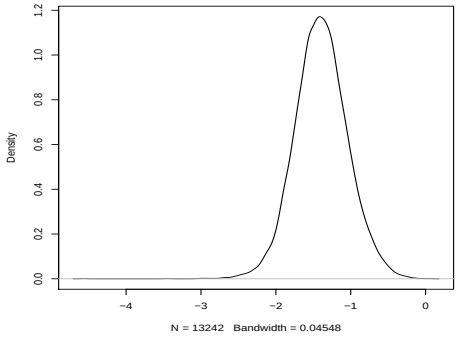


(l) Log phonotactic probability (biphone)

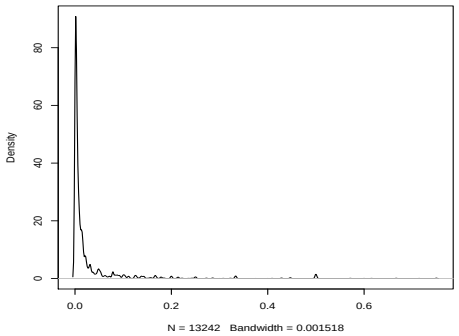
Figure 4.1: Continued: Distribution of type-based variables in the word duration model. Distributions are based on word types (n=540). For orthographic length, histogram is shown. For all other variables, probability functions are shown.



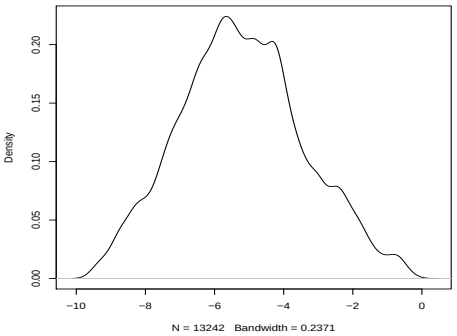
(a) Raw word token duration



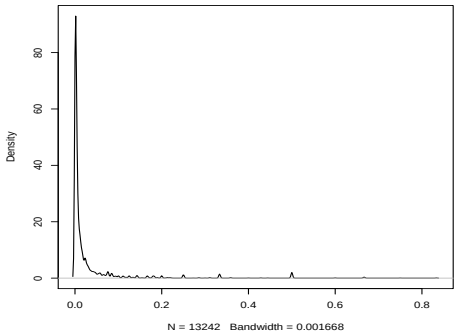
(b) Log word token duration



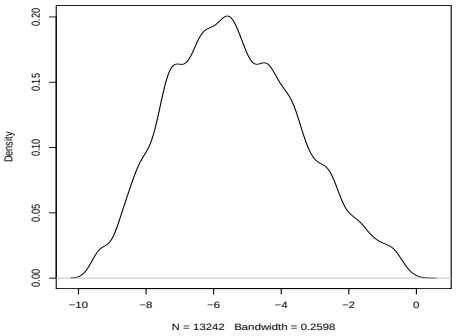
(c) Raw bigram probability (before)



(d) Log bigram probability (before)

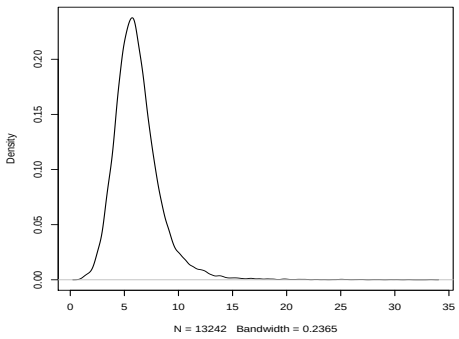


(e) Raw bigram probability (after)

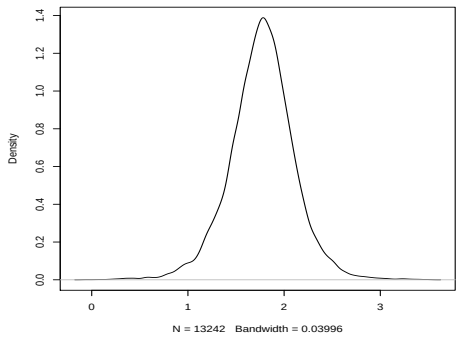


(f) Log bigram probability (after)

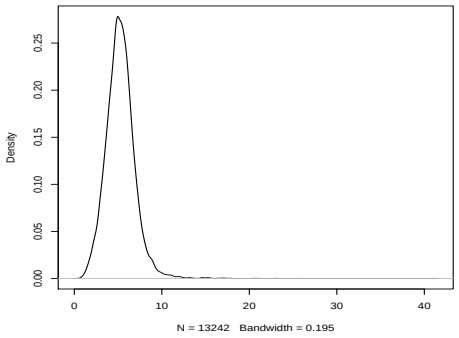
Figure 4.2: Distribution of token-based variables in the word duration model. For all variables, probability density functions over the token set (n=13,242) are shown. (Continued on the next page.)



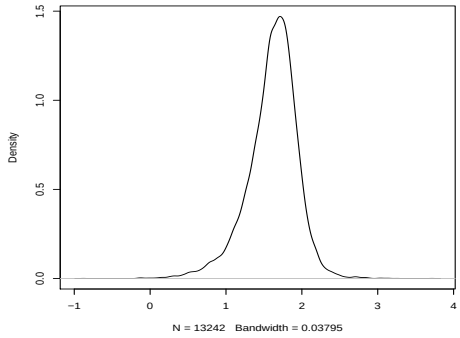
(g) Raw speech rate (before)



(h) Log speech rate (before)



(i) Raw speech rate (after)



(j) Log speech rate (after)

Figure 4.2: Continued: Distribution of token-based variables in the word duration model. For all variables, probability density functions over the token set (n=13,242) are shown.

Another crucial type of information for numerical variables is the correlation among variables. If two predictors are highly correlated, they may interfere with each other in the model. The next two tables show all pair-wise correlations among the variables. Again, type-based variables are separated from token-based variables, with type-based variables correlated based on the word set (Table 4.3) and token-based variables the token set (Table 4.4).

As shown in Table 4.3, neighborhood density is most highly correlated with phonotactic probability (r=0.62 with phoneme probability; r=0.57 with biphone probability). Not surprisingly, the two phonotactic measures are also highly correlated with each other (r=0.70). However, there is little correlation between neighbor frequency and neighborhood density (r=0.14), nor is there any sizable correlation between neighbor frequency and phonotactic probability (r≤0.2 with both phonotactic measures).

	ND	NFrq	BDur	Fam	Frq	OLen	PP_P	PP_Bi
ND	1	0.14	-0.10	-0.020	0.019	-0.29	0.62	0.57
NFrq	–	1	-0.15	-0.019	0.11	0.032	0.17	0.20
BDur	–	–	1	0.019	-0.062	0.21	-0.08	-0.15
Fam	–	–	–	1	0.25	-0.0091	-0.028	-0.007
Frq	–	–	–	–	1	0.075	-0.025	-0.058
OLen	–	–	–	–	–	1	-0.39	-0.39
PP_P	–	–	–	–	–	–	1	0.70
PP_Bi	–	–	–	–	–	–	–	1

Table 4.3: Pair-wise correlations among type-based numerical variables in the word duration model (ND = neighborhood density; NFrq = neighbor frequency; BDur = log baseline duration; Fam = familiarity; Frq = log frequency; OLen = orthographic length; PP_P = log phonotactic probability (phoneme); PP_Bi = log phonotactic probability (biphone)).

	WDur	Bigram_before	Bigram_after	Rate_before	Rate_after
WDur	1	-0.09	-0.26	-0.18	-0.22
Bigram_before	–	1	0.069	-0.018	0.025
Bigram_after	–	–	1	0.024	-0.063
Rate_before	–	–	–	1	0.16
Rate_after	–	–	–	–	1

Table 4.4: Pair-wise correlations among token-based numerical variables in the word duration model (WDur = log word token duration; Bigram_before = log bigram probability (before); Bigram_after = log bigram probability (after); Rate_before = log speech rate (before); Rate_after = log speech rate (after)).

4.3 Model

4.3.1 Model construction

A mixed-effect model is built to predict word token duration, with neighborhood density and neighbor frequency, as well as all the control variables as fixed-effects predictors. Speaker and word are entered as crossed random effects. All numerical predictors are centered (i.e. with the mean removed) before entering the model in order to reduce collinearity (see Baayen 2008 for a discussion on centering and collinearity).

As described in Chapter 3, I adopted a trimming procedure in model construction, so that insignificant predictors and outliers in the data would be inspected and if appropriate, eliminated. The two neighborhood measures, as well as the two phonotactic measures, are exempted from the trimming process, so that they have the maximal opportunity to account for variation in the model.

The specific steps of constructing the duration model are as follows. First, a model with all predictors was built. In the first round of elimination, three fixed-effects predictors (familiarity, disfluency (before), speaker sex) with absolute t values smaller than 1 were removed. Excluding these predictors did not cause a significant change in model fit, as shown by a likelihood ratio test (Chisq=0.26, Chi Df=3, p=0.96). A subsequent model was built with the remaining predictors. In the second round of elimination, all predictors with absolute t values under 2 were individually examined by means of model comparison. If omitting a predictor did not cause a significant change in model fit (p>0.05), the predictor would be considered as insignificant. Three variables (orthographic length, previous mention, speaker age) were examined and identified as insignificant. Removing these four predictors, a new model was built, which I will refer to as Model A.

The last step of the trimming procedure concerned potential outliers in the data. Figure 4.3 shows the distribution of the residuals in Model A against a normal distribution. The plot suggests that overall, the residuals are quite normally distributed, but there exists a small set of tokens with very large residuals. These are probably tokens that have extraordinarily short or long durations. To verify if this is the case, I divided the data set into two subsets: an "outlier" population (n=256) containing tokens with residuals greater than 2.5 standard deviations and a "normal" population with the rest of the tokens. Figure 4.4 shows that the distribution of raw word token duration in the "normal" population is basically the same as in the complete data set (cf. Figure 4.2), but in the "outlier" population, word duration follows a bimodal distribution, with one peak around 0.1s and the other around 0.5s. This suggests that the model is indeed having difficulty with predicting extreme duration values. If we remove the outliers from the data set and refit the model (Model B), the normality of the residuals will be greatly improved (see Figure 4.5).

To ascertain whether the existence of outliers could distort estimations of model coefficients, I compared the coefficients of Model A and Model B (see Table 4.5). For most predictors, the coefficient and associated statistics are unchanged in Model B. The coefficient of neighbor frequency changes sign (β=-0.012 in Model A; β=0.0029 in Model B), but the

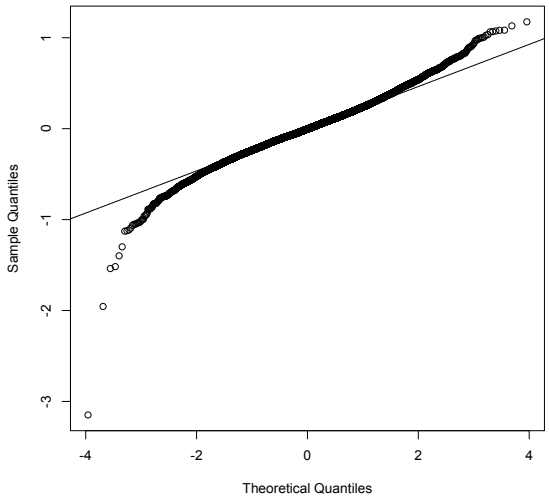


Figure 4.3: Q-Q plot of the residuals in Model A.

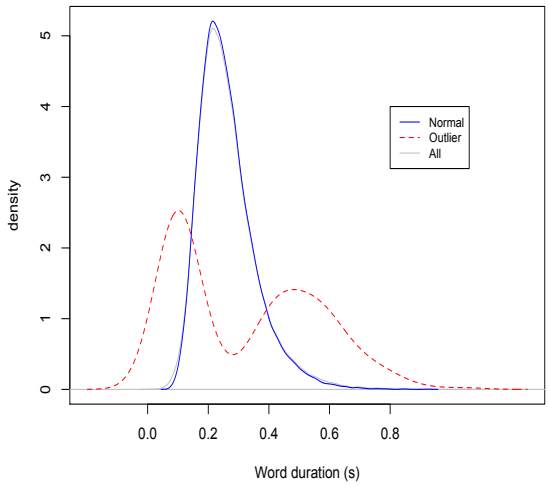


Figure 4.4: Distribution of word token duration in the "normal" population and the "outlier" population. Outliers are identified as data points associated with residuals greater than 2.5 standard deviations in Model A. Altogether there are 256 outliers.

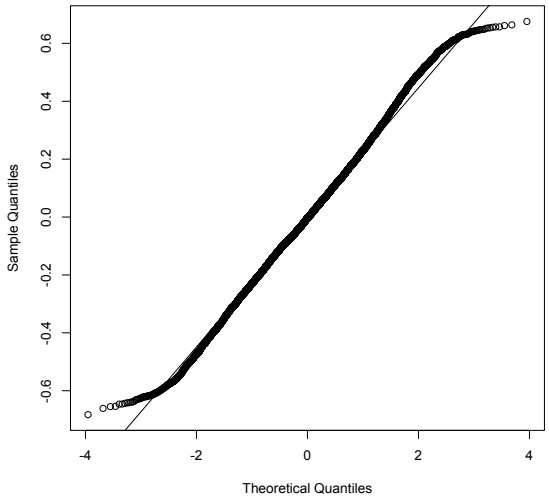


Figure 4.5: Q-Q plot of the residuals in Model B.

effect is not significant in either model ($|t|$<0.5). A possibly more significant change is seen in phoneme probability, for which the coefficient drops from 0.044 to 0.027 in Model B, accompanied by a decrease in t value from 1.63 to 1.05. This suggests that its weak significance (if any) in Model A is probably due to the presence of outliers.

In view of their association with extreme durations and possible distortion in some predictors, the 256 outliers were permanently removed from the model. The resulting model, i.e. Model B, contains 12,986 tokens from 539 word types and will be used as the basis for model discussion and evaluation in the rest of this chapter. Therefore, I will refer to this model as the baseline model hereafter[3].

4.3.2 Model summary

4.3.2.1 Model fit and random effects

The log likelihood of the baseline model is -66.97. With all random and fixed effects, the proportion of variance explained by the model (R^2) is 0.51. As many experimental studies have found, a large part of explained variance is accounted for by random variance among words and speakers. If we remove all fixed effects from the model, the model can still achieve an R^2 of 0.41. However, this does not mean that the inclusion of fixed-effects terms is

[3]It should be noted that the baseline model still contains non-significant predictors since neighborhood measures and phonotactic probability are exempt from the model trimming process. As a result, the estimates also reflect the model's attempt to incorporate insignificant predictors.

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.023	-60.14
neighborhood density	-0.0057	0.0011	-5.05
neighbor frequency	-0.012	0.026	-0.44
baseline duration	0.65	0.051	12.70
bigram probability (before)	-0.015	0.0015	-9.88
bigram probability (after)	-0.025	0.0014	-18.07
disfluency (after): T	0.36	0.011	33.19
frequency	-0.029	0.0045	-6.43
part of speech: ADV	-0.072	0.065	-1.11
part of speech: N	0.020	0.020	1.000
part of speech: V	-0.099	0.020	-4.91
phonotactic probability (phoneme)	0.044	0.027	1.63
phonotactic probability (biphone)	0.0034	0.013	0.26
speech rate (before)	-0.085	0.0072	-11.73
speech rate (after)	-0.13	0.0075	-17.55

(a) Model A: Before removing outliers

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.022	-60.71
neighborhood density	-0.0057	0.0011	-5.23
neighbor frequency	0.0029	0.025	0.11
baseline duration	0.68	0.050	13.66
bigram probability (before)	-0.014	0.0014	-10.54
bigram probability (after)	-0.023	0.0012	-18.79
disfluency (after): T	0.36	0.0099	36.57
frequency	-0.031	0.0044	-7.04
part of speech: ADV	-0.072	0.064	-1.13
part of speech: N	0.014	0.019	0.71
part of speech: V	-0.098	0.020	-5.01
phonotactic probability (phoneme)	0.027	0.026	1.05
phonotactic probability (biphone)	0.0026	0.013	0.200
speech rate (before)	-0.080	0.0066	-12.14
speech rate (after)	-0.13	0.0068	-19.53

(b) Model B: After removing outliers

Table 4.5: Summary of coefficients of the fixed-effects predictors in Model A (before removing outliers) and Model B (after removing outliers). Model A contains 13,242 tokens of 540 words. Model B contains 12,986 tokens of 539 word types. Outliers are identified as data points associated with residual greater than 2.5 standard deviations in Model A. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

trivial. As shown in Table 4.6, among other things, adding fixed effects reduces the standard deviation of by-word adjustment from 0.18 to 0.096 (i.e. 46% decrease) and the standard deviation of by-speaker adjustment from 0.11 to 0.092 (i.e. 16% decrease). The difference between word and speaker is expected, since all fixed effects in the model are linguistic factors (both speaker age and speaker sex have been eliminated in model construction).

Model	**Random effect**	**Variance**	**s.d.**
Baseline model	`word`	0.0093	0.096
	`speaker`	0.0084	0.092
	`residual`	0.056	0.24
No fixed-effects terms	`word`	0.032	0.18
	`speaker`	0.012	0.11
	`residual`	0.068	0.26

Table 4.6: Comparing random effects in the baseline model (upper) and a model with random effects only (lower).

4.3.2.2 Fixed effects

A potential problem in a mixed-effects model is collinearity, which is mainly caused by high correlations among numerical predictors. Overall collinearity of the model is usually assessed by a parameter called condition number, which can be generated by the `collin.fnc()` function in the **languageR** package (Baayen 2009). The condition number of the baseline model is extremely small (2.92), which suggests that there is no collinearity problem to be worried about.[4] Table 4.7 (repeated from Table 4.5) summarizes the fixed-effects terms in the baseline model. All predictors except for neighbor frequency and phonotactic probability (both phoneme and biphone) are associated with absolute t values greater than 2.5. As discussed in Chapter 3, a t value greater than 2.5 will roughly correspond to a p value smaller than 0.01, given the current data size. Thus I will rely on the rough estimation of significance levels of t for now. More sophisticated tests of t values will be discussed in Section 4.3.3 on model evaluation.

As shown in Table 4.7, neighborhood density is associated with a negative coefficient (-0.0057) with a relatively small standard error (0.0011) and a relatively large t value (-5.23). Taken together, the statistics suggest that a tendency exists for words from dense neighborhoods to be realized with *shorter* durations. Neighbor frequency, however, is associated with a small coefficient (0.0029) and a standard error that is ten times as big (0.025), resulting in a t value that is extremely small (0.11), which indicates no statistical significance.

Both phoneme probability and biphone probability are associated with positive coefficients (0.027 and 0.0026, respectively) and small t values (1.05 and 0.20, respectively).

[4]The small condition number is mostly due to centering. The condition number of uncentered numerical predictors is almost 90.

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.022	-60.71
neighborhood density	-0.0057	0.0011	-5.23
neighbor frequency	0.0029	0.025	0.11
baseline duration	0.68	0.050	13.66
bigram probability (before)	-0.014	0.0014	-10.54
bigram probability (after)	-0.023	0.0012	-18.79
disfluency (after): T	0.36	0.0099	36.57
frequency	-0.031	0.0044	-7.04
part of speech: ADV	-0.072	0.064	-1.13
part of speech: N	0.014	0.019	0.71
part of speech: V	-0.098	0.020	-5.01
phonotactic probability (phoneme)	0.027	0.026	1.05
phonotactic probability (biphone)	0.0026	0.013	0.200
speech rate (before)	-0.080	0.0066	-12.14
speech rate (after)	-0.13	0.0068	-19.53

Table 4.7: Summary of coefficients of the fixed-effects predictors in the baseline model. The model is based on 12,986 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

The rest of the fixed-effects predictors all seem to have strong effects on word duration, indicating that a word token is shorter if it is (a) composed by inherently short phonemes, (b) high in frequency, (c) a verb as opposed to other syntactic categories, (d) easily predictable from adjacent words, (e) not followed by disfluency, and (f) surrounded by fast speech. All these effects are in expected directions. The fact that following disfluency plays a more important role than preceding disfluency, and following speech rate has a larger size of effect than preceding speech rate (as shown in the magnitude of the coefficients), is in line with earlier observations that planning-related variation is often anticipatory (e.g. Shriberg 2001).

In addition to the significance and direction of an effect, we would also like to know how big the impact is (if it is significant) beyond other factors in the model. This can be done by examining the partial effect of a predictor, that is, the variation in the outcome variable when only the critical predictor is varying and all other predictors are controlled at the mean levels.

Figure 4.6 plots all partial effects in the baseline model, produced by the `plotLMER.fnc` function in the **languageR** package (Baayen 2009). The solid lines show the predicted partial effects and the broken lines show the MCMC-based confidence intervals (see the following section on model evaluation). As shown in the subplot for neighborhood density, there is a

clear negative relationship between density and word duration when other factors are controlled. The confidence intervals around the predicted line suggest that the relationship is quite linear throughout the range of the data. Overall, the partial effect of neighborhood density has a very similar pattern to that of bigram probabilities and frequency, all of which are well-established predictors for durational shortening. Moreover, based on the model parameters, we can also calculate the predicted range of variation in the outcome variable by varying neighborhood density while keeping all other predictors at their mean levels (henceforth "the predicted range of neighborhood density"). The predicted range of neighborhood density is [-1.47, -1.26] in log duration, which corresponds to [0.23s, 0.28s] in raw duration. For comparison, the predicted ranges (in raw duration) of preceding bigram probability, following bigram probability and frequency are [024s, 0.27s], [0.23s, 0.28s], [0.24s, 0.33s], respectively. This suggests that the size of the density effect on word duration is comparable to the size of the predictability effect but slightly smaller than the size of the frequency effect.

Not surprisingly, the partial effect of neighbor frequency is basically predicted to be a flat line, which indicates that there is no effect of neighbor frequency even when other predictors are kept constant. Therefore, there is also no size of the effect to speak of. The same pattern is observed with phoneme probability and biphone probability, both of which are associated with small t values.

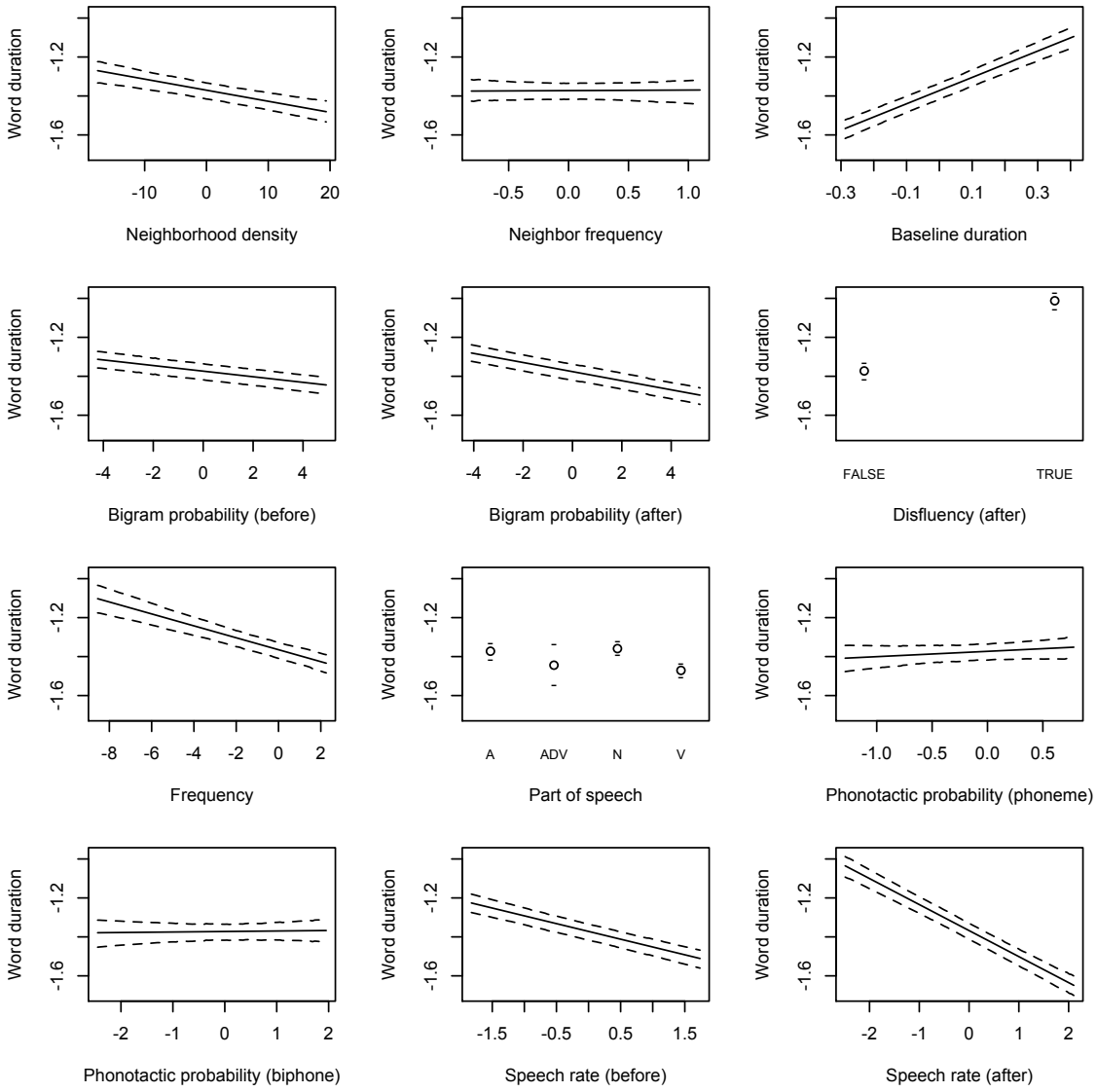


Figure 4.6: Partial effects in the baseline model. Y axis denotes log word duration in all subplots. X axis denotes individual predictor variable, after log transforming and centering, if any. The solid lines show the predicted lines for partial effects and the broken lines show the MCMC-base estimation of confidence intervals.

4.3.3 Model evaluation

Given the large size of the database and the subtlety of the effect under investigation, it is important to ensure that the observed effects are not spurious. A number of evaluation techniques (model comparison, MCMC-based testing and cross validation) were used to test the reliability and robustness of model results, especially the results regarding the neighborhood effects and phonotactic effects.

4.3.3.1 Comparing models with and without neighborhood measures

Model comparison is an effective way of gauging the significance of a particular predictor (or a set of predictors). In the current model, when both neighborhood density and neighbor frequency are removed, model fit deteriorates significantly (Chisq=27.29, Chi Df=2, $p<0.001$). However, if the two neighborhood measures are tested separately, it becomes clear that the change in model performance is mainly attributed to neighborhood density (Chisq=27.27, Chi Df=1, $p<0.001$), but not to neighbor frequency (Chisq=0, Chi Df=1, $p=1$). That is to say, neighborhood density, but not neighbor frequency, significantly improves the prediction for word duration.

On the other hand, removing the two phonotactic measures does not cause a significant change in model fit (Chisq=1.80, Chi Df=2, $p=0.41$), which confirms their insignificance.

4.3.3.2 Testing t values

A more straightforward indicator of statistical significance is the magnitude of the t value. However, as mentioned in Chapter 3, one problem with mixed-effect models is that it is not clear yet how to appropriately calculate degrees of freedom, and therefore it is hard to translate the t values into the more interpretable p values. A good way to go around the problem is to use Markov chain Monte Carlo (MCMC) sampling technique, which generates a large number of samples of model parameters so that an insight can be gained from the posterior distributions of the parameters.

Table 4.8 shows the summary statistics of MCMC-based estimation of model coefficients. Overall, MCMC coefficients coincide with those reported in model summary (cf. Table 4.7). In particular, for neighborhood density, the mean MCMC coefficient is -0.0057, which is exactly the same as the one given in Table 4.7. Moreover, the MCMC confidence interval of this parameter, [-0.0074, -0.0036], indicates that it is reliably smaller than zero ($p<0.0001$). Neighbor frequency, on the other hand, also has a mean MCMC coefficient that is the same as the one in Table 4.7 ($\beta=0.0029$), but the confidence interval, [-0.050, 0.039], covers the point zero almost in the middle, and therefore generates a large p value (>0.9), indicating lack of significance. The two phonotactic measures also have MCMC confidence intervals that cross zero ([-0.019, 0.070] for phoneme probability; [-0.019, 0.025] for biphone probability), and the associated p values are greater than 0.2 in both cases.

Thus the MCMC-based testing results confirm the significance of the neighborhood density effect as well as the insignificance of the neighbor frequency and phonotactic probability

effects.

Predictor	MCMC mean	HPD95 lower	HPD95 upper	Pr(>|t|)
(intercept)	-1.36	-1.41	-1.33	0.0000
neighborhood density	-0.0057	-0.0074	-0.0036	0.0000
neighbor frequency	0.0029	-0.050	0.039	0.9099
baseline duration	0.68	0.59	0.76	0.0000
bigram probability (before)	-0.014	-0.017	-0.012	0.0000
bigram probability (after)	-0.023	-0.026	-0.021	0.0000
disfluency (after): T	0.36	0.34	0.38	0.0000
frequency	-0.031	-0.039	-0.024	0.0000
part of speech: ADV	-0.072	-0.17	0.048	0.2606
part of speech: N	0.014	-0.017	0.048	0.4781
part of speech: V	-0.098	-0.13	-0.065	0.0000
phonotactic probability (phoneme)	0.027	-0.019	0.070	0.2945
phonotactic probability (biphone)	0.0026	-0.019	0.025	0.8406
speech rate (before)	-0.080	-0.093	-0.067	0.0000
speech rate (after)	-0.13	-0.15	-0.12	0.0000

Table 4.8: Summary of coefficients estimated by MCMC sampling technique for the baseline model. "MCMC mean" denotes the mean estimation of coefficients, based on 10,000 Monte Carlo samples of the posterior distribution of model parameters; "HPD95 lower" and "HPD95 upper" denote the lower and upper bounds of Highest Posterior Density interval for 95% of the probability density; "Pr (>|t|)" denotes the probability based on the t-distribution with 13,228 degrees of freedom.

4.3.3.3 Testing model generalizability

So far the significance of the neighborhood effects has been attested. A different question in model evaluation is how robust the effects are. After all, we are interested in the general effect of neighborhood structure on word duration, which goes beyond the current data set. To evaluate the robustness of the model results, I used two cross validation tests, a type-based one and a token-based one. The procedure and results of each test are presented below.

First, I tested whether the findings were an accident of the current word set. 100 samples were randomly drawn from the target word set (n=539), each containing about half of the target words. A test model was built for each sample, with all tokens of the sampled words. Model coefficients are summarized over the 100 test models (see Table 4.9). According to the summary table, neighborhood density has a mean coefficient of -0.0061 over all models, which is only slightly different than the one reported in model summary and MCMC estimation

(-0.0057). Importantly, in 95% of the test models, the coefficient is within [-0.0078, -0.0039], indicating that it is reliably different from zero. On the other hand, neighbor frequency has an average coefficient of 0.0064, but its confidence interval, [-0.048, 0.046], spans across the point of zero, suggesting a lack of significance. The same pattern is shown in phonotactic measures. Taken together, type-based cross validation results converge with findings from the baseline model, which suggests that the observed effects (or lack of effect) persist across subsets of words and are not likely to be an artifact of the current word set.

Second, I used the same method to test the effect of tokens. This time, 100 random samples were drawn directly from the token set (n=12,986), with no constraint on word types. Each sample ontains about half of the observations. 100 test models were constructed based on the token samples. Model coefficients are summarized over all test models (Table 4.9). The results converge again with findings from the baseline model. Neighborhood density has a mean coefficient of -0.0054 with a confidence interval of [-0.0063, -0.0044], indicating statistical significance; however, neighbor frequency, as well as the two phonotactic measures, are estimated with coefficients that are not reliably different from zero. These results suggest that findings from the baseline model also persist across subsets of the tokens, and therefore are not likely to be an artifact of the current token set.

Taken together, cross validation results confirm that the observed effects in the baseline model may be generalized to unseen words and tokens that are similar in nature to the current data.

4.3.3.4 Summary of model evaluation results

To sum up, the baseline model shows a strong negative effect of neighborhood density on word duration and null effects of neighbor frequency and phonotactic probability. The reliability and robustness of these findings is confirmed by all evaluation tests that have been applied. Everything else being equal, a high density word tends to be shorter than a low density word, and the size of the effect is comparable to that of word predictability.

Predictor	2.5%	50%	97.5%
(intercept)	-1.39	-1.37	-1.31
neighborhood density	-0.0078	-0.0061	-0.0039
neighbor frequency	-0.048	0.0064	0.046
baseline duration	0.59	0.67	0.76
bigram probability (before)	-0.023	-0.015	-0.011
bigram probability (after)	-0.027	-0.023	0.35
disfluency (after): T	-0.071	0.36	0.400
frequency	-0.039	-0.030	-0.023
part of speech: ADV	-0.200	-0.067	0.089
part of speech: N	-0.095	0.010	0.044
part of speech: V	-0.13	-0.098	0.012
phonotactic probability (phoneme)	-0.026	0.033	0.080
phonotactic probability (biphone)	-0.021	0.0010	0.026
speech rate (before)	-0.13	-0.083	-0.066
speech rate (after)	-1.37	-0.14	-0.12

(a) Type-based cross validation

Predictor	2.5%	50%	97.5%
(intercept)	-1.39	-1.37	-1.35
neighborhood density	-0.0063	-0.0054	-0.0044
neighbor frequency	-0.040	-0.0099	0.0098
baseline duration	0.61	0.68	0.73
bigram probability (before)	-0.017	-0.015	-0.011
bigram probability (after)	-0.026	-0.024	-0.022
disfluency (after): T	0.34	0.36	0.38
frequency	-0.036	-0.031	-0.026
part of speech: ADV	-0.11	-0.057	-0.0016
part of speech: N	0.0025	0.020	0.037
part of speech: V	-0.11	-0.097	-0.077
phonotactic probability (phoneme)	-0.0039	0.020	0.046
phonotactic probability (biphone)	-0.011	0.0043	0.017
speech rate (before)	-0.095	-0.082	-0.071
speech rate (after)	-0.15	-0.14	-0.12

(b) Token-based cross validation

Table 4.9: Summary of cross validation results for the baseline model. Type-based cross validation is based on 100 random subsets of half of the word types; token-based cross validation is based on 100 random subsets of half of the tokens. In both tables, "50%" denotes the mean value of the coefficient over 100 models; "2.5%" and "97.5%" denote the lower and upper bounds of the 95% density area of the coefficient.

4.4 Alternative analyses

A remaining concern with the current results regards the calculation of neighborhood metrics. Is it possible that the observed effect is an artifact of how neighborhood metrics are coded now? In this section, I will address this concern by presenting a series of models with alternative neighborhood metrics.

As reviewed in Chapter 2, the literature on phonological neighborhoods has exploited various ways of measuring neighborhood density. For one thing, many studies have used a combined measure of density and neighbor frequency (i.e. frequency-weighted neighborhood density). The underlying assumption is that neighbor frequency operates in a similar way as neighborhood density: a high frequency neighbor exerts a large influence on the target word, just like a dense neighborhood. In the current study, density and neighbor frequency are kept distinct (up to this point) intentionally, in order to separate the effects of these two factors. However, for the purpose of both assessing model reliability and comparing with earlier results, it makes sense to test the current data with a frequency-weighted density measure.

Another line of deviation concerns the dictionary for calculating neighborhoods. The current measures are from the Hoosier mental lexicon, which is by far the most often cited reference in the neighborhood density literature. However, there are two potential problems with the HML measures. First of all, the identification of phonological neighbors in HML does not consider word frequency or familiarity. Thus, the estimated neighborhood may contain words that are not present (or actively present) in speakers' mental lexicon. Second, the frequency measure in HML comes from Kučera and Francis (1967), which may not be an accurate estimate for usage frequency in English in the late nineties, when the Buckeye corpus was recorded. In fact, Brysbaert and New (2009) has shown that the K&F frequency is less good at predicting psycholinguistic norming data compared with some other more recent corpus-based word frequencies (e.g. CELEX, HAL, SUBTLEX, etc). Thus it seems necessary to test the model with at least one other set of neighborhood metrics compiled from a different dictionary. In the current work, I compiled an alternative set of neighborhood metrics, using word frequency from the CELEX database and incorporating a frequency threshold for neighborhood membership. The CELEX database was chosen over other more recent and accurate frequency dictionaries (e.g. SUBTLEX, according to Brysbaert & New) because it is a more comprehensive lexical database and some of the lexical information it provides (i.e. part of speech) is already used in the current work. Furthermore, CELEX has been explored in previous research on neighborhood density (Frauenfelder et al. 1993). The set of neighborhood metrics compiled from CELEX will be referred to as the "CELEX-based" neighborhood metrics.

Although there are other ways that neighborhoods can be measured, e.g., log transformed neighborhood density (though it is already normally distributed), maximal and minimal neighbor frequency, etc, in this work, I only pursued the two alternative neighborhood measures just discussed, that is, frequency-weighted neighborhood density and CELEX-based neighborhood metrics. In the rest of this section, I will present the results from these

two alternative analyses. All the models were built with the 13,242-token database described in Chapter 3. Model construction followed the same trimming procedure as presented in Section 4.3.1, so that the best model performance could be achieved. As a result, the structure of the alternative models may deviate from that of the baseline model. For each alternative model, a summary of fixed effects will be reported, as well as selective findings from model evaluation. To preview the results, all the alternative models suggest the same trend as seen in the baseline model, that is, words from dense neighborhood are realized with shorter durations.

4.4.1 Using frequency-weighted neighborhood density

4.4.1.1 Calculating frequency-weighted neighborhood density

The formula of frequency-weighted neighborhood density (FWND) can be written as $\sum f_i \cdot 1$, where f_i is the frequency of individual neighbor. Thus FWND equals the sum of neighbor frequency and can be conveniently derived by multiplying (average) neighbor frequency with neighborhood density (i.e. $\bar{f} \cdot ND$), both of which already exist.

FWND in the data set ranges from 4.43 to 91.99 (in log frequency), with a median value of 43.66 and a mean value of 45.01 (s.d. = 17.59). Figure 4.7 plots the distribution of FWND over word types. Interestingly, FWND is much more correlated with raw density (r=0.95) than with neighbor frequency (r=0.41)[5].

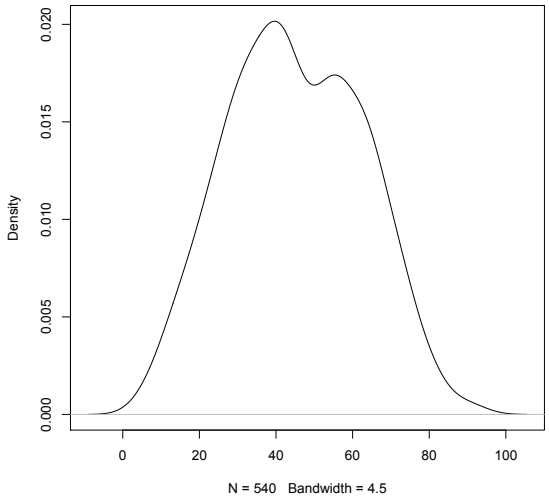


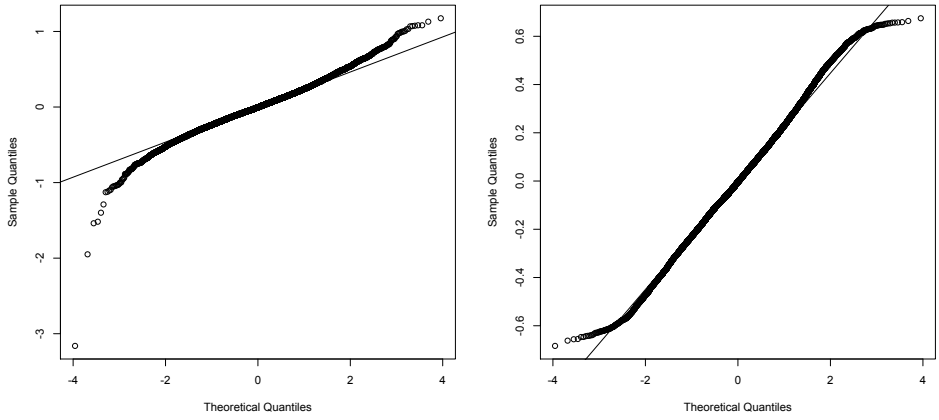
Figure 4.7: Distribution of FWND in the word duration database. The probability density function is based on word types (n=540).

[5]It is not clear yet if the high correlation between raw and weighted density is an accident of the current data or a general fact of the language.

4.4.1.2 Modeling with frequency-weighted neighborhood density

First an initial model was built based on the 13,242-token database, with FWND and all the control factors. In the trimming process, 5 predictors were removed (first round: disfluency (before), familiarity, speaker sex; second round: previous mention, speaker age), as well as 257 outlier data points. The removal of outliers significantly improves model fit (see Figure 4.8). Interestingly, the effect of phoneme probability is again, reduced when outliers are removed, with β changing from 0.053 to 0.037 and the associated t value from 1.94 to 1.42 (see Table 4.10). This supports the previous speculation that the marginal significance of phoneme probability is mostly due to the presence of outliers.

The final model with FWND (i.e. the FWND model) contains 12 fixed-effects terms and 2 random terms, based on 12,985 tokens of 539 words.



(a) Before removing outliers (b) After removing outliers

Figure 4.8: Q-Q plot of the residuals in the FWND Model, before and after removing 257 outliers.

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.023	-60.20
FWND	-0.0026	0.00050	-5.18
baseline duration	0.62	0.052	11.83
bigram probability (before)	-0.015	0.0015	-9.86
bigram probability (after)	-0.025	0.0014	-18.07
disfluency (after): T	0.36	0.011	33.21
frequency	-0.029	0.0045	-6.33
orthographic length	0.020	0.0099	2.01
part of speech: ADV	-0.073	0.064	-1.13
part of speech: N	0.019	0.020	0.97
part of speech: V	-0.100	0.020	-4.99
phonotactic probability (phoneme)	0.053	0.027	1.94
phonotactic probability (biphone)	0.0094	0.013	0.700
speech rate (before)	-0.085	0.0072	-11.73
speech rate (after)	-0.13	0.0075	-17.54

(a) Before removing outliers

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.022	-60.87
FWND	-0.0025	0.00048	-5.20
baseline duration	0.64	0.050	12.70
bigram probability (before)	-0.014	0.0014	-10.43
bigram probability (after)	-0.023	0.0012	-18.81
disfluency (after): T	0.36	0.0098	36.63
frequency	-0.030	0.0043	-6.94
orthographic length	0.021	0.0096	2.21
part of speech: ADV	-0.073	0.063	-1.15
part of speech: N	0.012	0.019	0.63
part of speech: V	-0.10	0.020	-5.12
phonotactic probability (phoneme)	0.037	0.026	1.42
phonotactic probability (biphone)	0.0093	0.013	0.72
speech rate (before)	-0.080	0.0066	-12.11
speech rate (after)	-0.13	0.0068	-19.54

(b) After removing outliers

Table 4.10: Summary of coefficients of the fixed-effects predictors for the FWND model before and after removing outliers. The original model contains 13,242 tokens of 540 words. After removing outliers, the model contains 12,985 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

Table 4.10 suggests that FWND has a negative effect on word duration (β=-0.0025, t=-5.20 in the trimmed dataset, see 4.10(b)). Phonotactic probability, either phoneme or biphone, fails to reach significance ($|t|$<1.5 in both cases). Other fixed effects, including the intercept, are largely the same as in the baseline model. The only "new" predictor is orthographic length, which is eliminated from the baseline model but seems to be marginally significant in the FWND model (β=0.021, std. error=0.0096, t=2.21), suggesting that CVC words with longer spelling forms tend to be longer in duration.

Significance and robustness of the FWND effect is confirmed by model comparison (Chisq=26.958, Chi Df=1, p< 0.001), MCMC-based evaluation and cross validation. Results from the latter two tests are selectively reported in Table 4.11. The predicted range (in raw duration) of FWND is [0.23s, 0.28s], which coincides with that of raw density in the baseline model. As shown in Table 4.11, the insignificance of phonotactic probability (phoneme and biphone) is also confirmed by model evaluation.

To sum up, the FWND model shows that a combined measure of density and neighbor frequency produces a similar effect on word duration as raw density in the baseline model. This is not surprising given the high correlation between FWND and raw density, but importantly, it suggests that findings from the baseline model can be compared with previous results with FWND, regardless of whether the correlation is a coincidence of the particular set of words included in this study. Furthermore, the FWND model makes it possible to compare current results with findings from previous research that also used frequency-weighted neighborhood measures. I will return to this point in Chapter 6.

Predictor	MCMC mean	HPD95 lower	HPD95 upper	Pr(>|t|)
FWND	-0.0025	-0.0033	-0.0016	0.0000
phonotactic probability (phoneme)	0.037	-0.0099	0.082	0.1563
phonotactic probability (biphone)	0.0093	-0.013	0.032	0.4729

(a) MCMC-based evaluation

Predictor	2.5%	50%	97.5%
FWND	-0.0025	-0.0025	-0.0033
phonotactic probability (phoneme)	0.037	0.035	-0.0099
phonotactic probability (biphone)	0.0093	0.0091	-0.013

(b) Type-based cross validation

Predictor	2.5%	50%	97.5%
FWND	-0.0029	-0.0024	-0.0019
phonotactic probability (phoneme)	0.0072	0.029	0.052
phonotactic probability (biphone)	-0.0032	0.0089	0.022

(c) Token-based cross validation

Table 4.11: Summary of model evaluation results for the FWND model. MCMC-based evaluation was based on 10,000 Monte Carlo samples of posterior distribution of the model coefficients. Type-based cross validation was conducted with 100 random subsets of the data, each containing half of the word types. Reported model coefficients are summarized over all test samples. Token-based cross validation was based on 100 random subsets of the data, each containing half of the word tokens. Reported model coefficients are summarized over all test samples. Only results regarding FWND and phonotactic probability are shown in the table.

4.4.2 Using CELEX-based neighborhood measures

4.4.2.1 Calculating CELEX-based neighborhood measures

The CELEX lexical database (Baayen et al. 1993) was jointly developed by a number of research institutions in the nineties. CELEX frequency for English is based on the 18 million-word COBUILD corpus (Sinclair 1987), covering about 90,000 orthographic forms. For the current analysis, three CELEX-based neighborhood measures (neighborhood density, neighbor frequency and FWND) were coded. The coding procedure is described below.

First , for each target word, a set of potential neighbors were compiled based on pronunciation similarity. Since CELEX transcriptions feature British English instead of American English pronunciation, I used the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) instead for determining phonological similarity. The CMU dictionary is a public-domain pronouncing dictionary created by the Carnegie Mellon University for American English. Any word associated with a CMU pronunciation that is one phoneme away from the target word's CMU pronunciation by addition, deletion, or substitution of a phoneme would be considered as a potential neighbor.[6]

Next, the set of candidate neighbors were further screened for whether or not they should be considered as familiar to the speaker. Vitevitch (2002) used familiarity of 6 or higher (on a 7-point scale; Nusbaum et al. 1984) as the criterion for neighborhood membership. In the current work, since CELEX is much bigger than any existing psycholinguistic database, it is not possible to get familiarity ratings for all the words that need to be screened. Therefore I approximated familiarity with word frequency. Though the two measures cannot be equated (Colombo, Pasini, and Balota 2006), it is widely agreed that they are closely related (e.g. Balota and Chumbley 1985). High-frequency words tend to have high familiarity and highly familiar words tend to be frequently used.

Then the problem is to determine a reasonable threshold in frequency that would be more or less equivalent to a familiarity rating of 6. To gain some insight into the relationship between the two measures, I compared word frequency in the CELEX wordform dictionary with familiarity ratings from the HML dictionary, for words that exist in both (n=11,238). Figure 4.9 shows the distribution of log CELEX form frequency (i.e. $ln(Freq+1)$) of words with familiarity ratings between 4 and 5 (n=935), 5 and 6 (n=1264), and 6 and 7 (n=3897) in HML. As shown in the plot, words with higher familiarity are in general higher in frequency, but there is a great proportion of overlap, especially among words with lower familiarity. The mean log frequencies of the three groups are 2.40, 2.78, and 4.50, respectively (from low to high familiarity). In practice, I tested with both a lower threshold (log frequency=2.5) and a higher threshold (log frequency=4) for neighborhood membership, and the model results were similar. Therefore, I arbitrarily choose to report the results associated with the higher threshold. Under this constraint, only words with a log CELEX frequency of 4 or higher can

[6]Thus the set of potential neighbors is also limited by word forms listed in the CMU dictionary. This should not be a problem because the CMU dictionary is based on word forms and quite large in size (≈125,000).

Variable	Min	Max	Median	Mean	s.d.
CELEX density	1	36	18	17.34	5.89
CELEX neighbor frequency	5.07	8.37	6.47	6.57	0.55
CELEX FWND	7.22	245.40	116.70	113.50	39.04

Table 4.12: Summary statistics for CELEX density, CELEX neighbor frequency and CELEX FWND in the database, based on 13,242 tokens.

be considered as a neighbor. Words that meet the criterion have at least 54 occurrences in the 18 million-word COBUILD corpus.

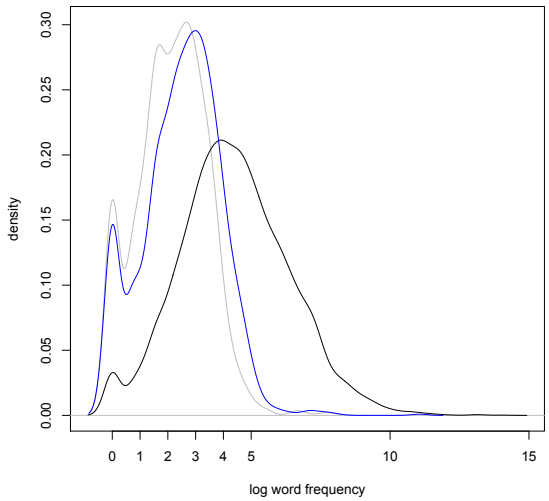
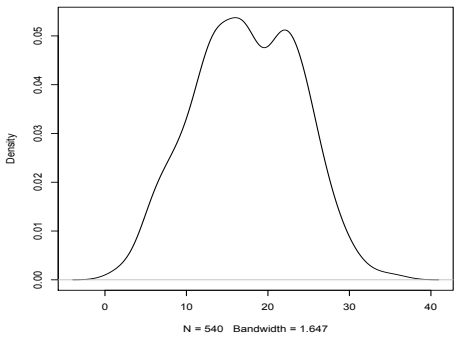
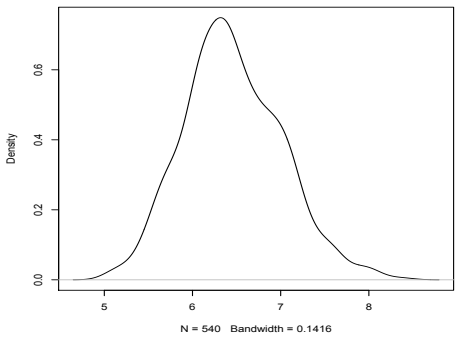


Figure 4.9: Distribution of CELEX log frequency (i.e. $ln(Freq+1)$) in words with familiarity ratings between 4 and 5 (grey), 5 and 6 (blue), and 6 and 7 (black) in HML. Word counts for the three groups are 935, 1264, 3897, correspondingly.

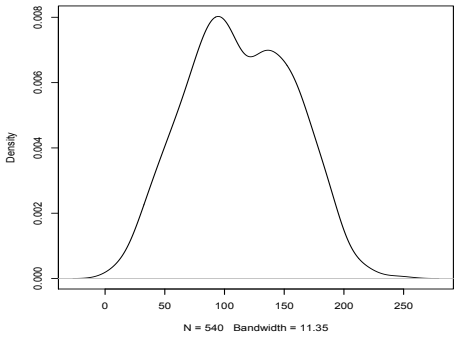
Thus, the neighborhood of each target word was determined. Three neighborhood metrics were calculated: raw density ("CELEX density"), mean log neighbor frequency ("CELEX neighbor frequency"), and log frequency-weighted density ("CELEX FWND"). Summary statistics of these variables are given in Table 4.12. Figure 4.10 plots the distribution of the variables over the word set. Both CELEX density and CELEX neighbor frequency are correlated with the corresponding HML metric (r=0.81 for density; r=0.65 for neighbor frequency). Similar to the HML metrics, CELEX FWND is extremely highly correlated with CELEX density (r=0.98), and sizably correlated with CELEX neighbor frequency (r=0.38).



(a) CELEX density



(b) CELEX neighbor frequency

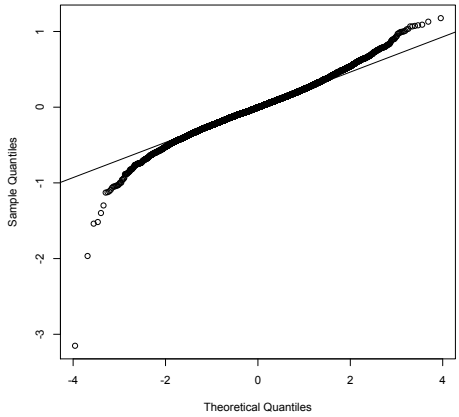


(c) CELEX FWND

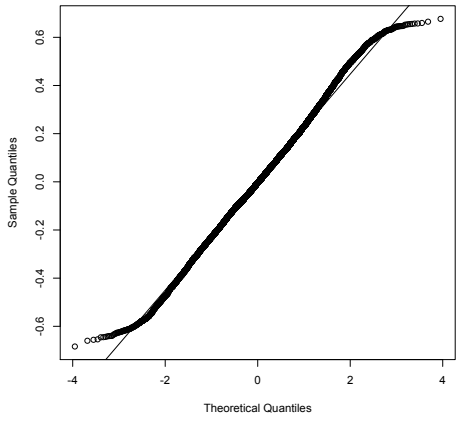
Figure 4.10: Distribution of CELEX-based neighborhood measures in the duration database. The probability density function is based on word types (n=540).

4.4.2.2 Modeling with CELEX-based neighborhood measures

Model construction was carried out with the same procedure as described before. Six variables were eliminated during the trimming process (first round: disfluency (before), familiarity, speaker sex; second round: orthographic length, previous mention, speaker age). In addition, 257 outliers were removed from the data. Figure 4.11 shows the distribution of model residuals before and after the removal. Table 4.13 summarizes model coefficients before and after the removal. The final model (i.e. the CELEX model) contained 12 fixed-effects predictors, based on 12,985 tokens and 539 words.



(a) Before removing outliers



(b) After removing outliers

Figure 4.11: Q-Q plot of the residuals in the CELEX model before and after removing 257 outliers.

Predictor	β	Std. Error	t value
(intercept)	-1.37	0.023	-59.79
CELEX density	-0.0037	0.0013	-2.91
CEKEX neighbor frequency	-0.0037	0.012	-0.300
baseline duration	0.64	0.054	11.89
bigram probability (before)	-0.015	0.0015	-9.90
bigram probability (after)	-0.025	0.0014	-18.07
disfluency (after): T	0.36	0.011	33.23
frequency	-0.028	0.0046	-6.06
part of speech: ADV	-0.066	0.066	-1.000
part of speech: N	0.024	0.020	1.19
part of speech: V	-0.096	0.021	-4.68
phonotactic probability (phoneme)	0.0071	0.026	0.27
phonotactic probability (biphone)	-0.0034	0.014	-0.25
speech rate (before)	-0.085	0.0072	-11.73
speech rate (after)	-0.13	0.0075	-17.52

(a) Before removing outliers

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.023	-60.44
CELEX density	-0.0039	0.0012	-3.24
CELEX neighbor frequency	-0.00061	0.012	-0.050
baseline duration	0.66	0.052	12.80
bigram probability (before)	-0.014	0.0014	-10.55
bigram probability (after)	-0.023	0.0012	-18.83
disfluency (after): T	0.36	0.0099	36.63
frequency	-0.030	0.0044	-6.68
part of speech: ADV	-0.068	0.065	-1.05
part of speech: N	0.015	0.019	0.79
part of speech: V	-0.097	0.020	-4.85
phonotactic probability (phoneme)	-0.0084	0.025	-0.34
phonotactic probability (biphone)	-0.0012	0.013	-0.10
speech rate (before)	-0.079	0.0066	-12.05
speech rate (after)	-0.13	0.0068	-19.51

(b) After removing outliers

Table 4.13: Summary of coefficients of the fixed-effects predictors for the CELEX model before and after removing outliers. The original model contains 13,242 tokens of 540 word types. After removing outliers, the model contains 12,985 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

Interestingly, though the neighborhood measures are calculated from different sources, the CELEX model reveals highly similar results compared with previous models. Neighborhood density is estimated with a negative coefficient (-0.0039) and a relatively large t value (-3.24), whereas neighbor frequency fails to reach significance (β = -0.00061, std. error = 0.012, t = -0.050). Phonotactic probability (either phoneme or biphone) fails to reach significance ($|t|$<0.5 in both cases).

The significance of the density effect is confirmed by model comparison (Chisq = 10.61, Chi Df = 1, p<0.001), MCMC-based testing and cross validation. Results from the latter two tests are summarized in Table 4.14. The predicted range (in raw duration) of CELEX density is [0.24s, 0.28s], which is similar to that of HML density in the baseline model (i.e. [0.23s, 0.28s]).

Predictor	MCMC mean	HPD95 lower	HPD95 upper	Pr(>|t|)
CELEX density	-0.0039	-0.0059	-0.0017	0.0012
CELEX neighbor frequency	-0.00060	-0.024	0.017	0.9592
phonotactic probability (phoneme)	-0.0084	-0.053	0.034	0.7374
phonotactic probability (biphone)	-0.0012	-0.023	0.021	0.9242

(a) MCMC-based evaluation

Predictor	2.5%	50%	97.5%
CELEX density	-0.0069	-0.0040	-0.0023
CELEX neighbor frequency	-0.021	0.00006	0.024
phonotactic probability (phoneme)	-0.041	-0.0061	0.041
phonotactic probability (biphone)	-0.027	-0.0042	0.023

(b) Type-based cross validation

Predictor	2.5%	50%	97.5%
CELEX density	-0.0050	-0.0038	-0.0027
CELEX neighbor frequency	-0.016	-0.0053	0.0075
phonotactic probability (phoneme)	-0.033	-0.012	0.014
phonotactic probability (biphone)	-0.012	-0.00013	0.012

(c) Token-based cross validation

Table 4.14: Summary of model evaluation results for the CELEX model. MCMC-based evaluation was based on 10,000 Monte Carlo samples of posterior distribution of the model coefficients. Type-based cross validation was conducted with 100 random subsets of the data, each containing half of the word types. Reported model coefficients are summarized over all test samples. Token-based cross validation was based on 100 random subsets of the data, each containing half of the word tokens. Reported model coefficients are summarized over all test samples. Only results regarding CELEX neighborhood measures and phonotactic probability are shown in the table.

Finally, given the high correlation between CELEX density and CELEX FWND (r = 0.98), it should be expected that CELEX FWND will behave just like CELEX density, and this is indeed the case. For conciseness, I will only show a summary of model coefficients with CELEX FWND (Table 4.15), which should be sufficient, given the consistent patterns that have been observed so far.

To sum up, results from the CELEX model corroborate findings from the baseline model, suggesting that the observed density effect is not sensitive to some fluctuation in the calculation method. Meanwhile, the insignificance of neighbor frequency also persists across models and therefore cannot be ascribed to the goodness of the frequency measure.

Predictor	β	**Std. Error**	t **value**
(intercept)	-1.37	0.023	-60.63
CELEX FWND	-0.00056	0.00017	-3.24
baseline duration	0.65	0.050	13.02
bigram probability (before)	-0.014	0.0014	-10.55
bigram probability (after)	-0.023	0.0012	-18.85
disfluency (after): T	0.36	0.0099	36.64
frequency	-0.029	0.0044	-6.64
part of speech: ADV	-0.069	0.064	-1.08
part of speech: N	0.016	0.019	0.85
part of speech: V	-0.096	0.020	-4.84
phonotactic probability (phoneme)	-0.012	0.025	-0.48
phonotactic probability (biphone)	-0.00049	0.013	-0.040
speech rate (before)	-0.079	0.0066	-12.06
speech rate (after)	-0.13	0.0068	-19.50

Table 4.15: Summary of coefficients of the fixed-effects predictors for the model with CELEX-based FWND measure. The model is based on 12,985 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

4.5 Chapter discussion

4.5.1 Potential confounding factors

In previous sections, I have presented findings from a major model (i.e. the baseline model) and a number of alternative models. All the models have consistently shown a negative effect of neighborhood density on word duration. Everything else being equal, words from dense neighborhoods are realized with shorter durations than words from a sparse neighborhoods. The significance of the density effect is further attested by model comparison and MCMC-based evaluation. Cross validation results suggest that the effect persists across subsets of words and tokens, and therefore is not likely to be an artifact of the current data set. Furthermore, results from alternative models indicate that the effect is robust enough to tolerate several types of deviation in neighborhood calculation (raw or frequency-weighted density, HML or CELEX frequency, with or without frequency threshold for neighborhood membership). In addition, model results also suggest that the size of the density effect is comparable to that of the predictability effect but smaller than the frequency effect.

However, one may still ask, *is it possible that the observed density effect is at least partially confounded by phonotactic probability?* Before we can claim a true density effect , it is necessary to fully discuss the relationship between neighborhood density and phonotactic probability in the model.

We know that by default, words from dense neighborhoods contain high frequency phonological sequences (Vitevitch et al. 1999), and the correlation is also manifested in the current database. Moreover, phonotactic probability by itself has been associated with facilitation in perception and production as well as durational shortening in elicited speech (e.g. Frisch, Large, and Pisoni 2000; Munson 2001; Vitevitch et al. 1999) and spontaneous speech (Van Son and Pols 2003), which makes it an even more likely confounding factor for the observed density effect. Nevertheless, in most of the models that have been investigated so far, phonotactic probability fails to show any significant effect whatsoever, even though it is always present in the model. In a couple models where phonotactic probability seemed to be marginally significant, the significance was later found out to be spurious and attributable to outliers in the data. Thus, in the very least, we can confirm that neighborhood density is a stronger and more effective predictor for word duration than phonotactic probability is.

Next, we would like to know if the insignificance of phonotactic probability is related with the presence of neighborhood density. To examine this possibility, I build a variant of the baseline model, which includes all the predictors except for neighborhood density and neighbor frequency. As shown in Table 4.16, the model shows that neither of the phonotactic variables reaches significance ($|t|<1.5$ in both cases). But this may also be due to the interference between the two phonotactic measures since they are also correlated (r=0.7). Therefore I further the test by removing one of the phonotactic variables (while keeping neighborhood measures absent). This time, the results show that with only one phonotactic variable in the model, either phoneme or biphone, and no neighborhood measures present,

the phonotactic effect seems to have an increased significance in the model. Table 4.17 summarizes the coefficients of the two relevant model variants. As shown in the table, if only phoneme probability is in the model, it is associated with a coefficient of -0.046 and a t value of -2.6; if only biphone probability is in the model, it is associated with a coefficient of -0.025 and a t value of -2.76. In both models, the predicted range (in raw duration) of the phonotactic variable is around [0.25s, 0.27s], which is slightly smaller than that of neighborhood density in the baseline model ([0.23s, 0.28s]).[7] Taken together, these results suggest that regardless of the competition with neighborhood density, the two phonotactic variables are also competing with each other and as a result, they suppress each other in the model.

This makes one wonder whether the insignificance of both phonotactic variables in the baseline model is also due to their copresence instead of the competition from neighborhood density. To examine this possibility, I build another pair of model variants based on the baseline model. This time, only one of the phonotactic variables is present in each model, together with neighborhood variables and other control variables (see Table 4.18). In these two models, neighborhood density consistently shows a strong negative effect ($\beta \approx$ -0.0055, std. error ≈ 0.001, $t \approx$ -5), with the same predicted range as in the baseline model. However, neither phoneme probability or biphone probability turns out to be significant ($|t|<1.5$ in both cases). These results indicate that both phoneme probability and biphone probability are also suppressed by neighborhood density.

Finally, to complete the analysis, I also test the possibility that the effect of neighborhood density is somehow dependent on the presence of at least one phonotactic variable. A model variant is built based on the baseline model, with all predictors but not the two phonotactic measures (Table 4.19). In this model, neighborhood density is still significant but with a slightly reduced coefficient compared to the baseline model(β=-0.0048, std. error=0.00082, t=-5.8).

In summary, the current evidence suggests that both phonotactic variables are indeed overshadowed by neighborhood density in the model, and they also tend to suppress each other when both are present. However, the effect of neighborhood density exists regardless of the presence or absence of either or both of the phonotactic variables. In addition, when compared independently, the density effect is also larger in size than the phonotactic effect. Taken together, part of the density effect on word duration may also be explained by phonotactic probability, but phonotactics alone cannot fully account for the density effect.

[7] It should be noted that these models are all constructed based on the data set of the baseline model, which has already excluded outliers that tend to have extreme durations.

Predictor	β	**Std. Error**	t **value**
(intercept)	-1.37	0.023	-60.28
baseline duration	0.68	0.050	13.41
bigram probability (before)	-0.014	0.0014	-10.60
bigram probability (after)	-0.023	0.0012	-18.77
disfluency (after): T	0.36	0.0099	36.64
frequency	-0.031	0.0044	-6.98
part of speech: ADV	-0.055	0.066	-0.84
part of speech: N	0.020	0.019	1.03
part of speech: V	-0.098	0.020	-4.88
phonotactic probability (phoneme)	-0.023	0.025	-0.94
phonotactic probability (biphone)	-0.016	0.012	-1.31
speech rate (before)	-0.080	0.0066	-12.13
speech rate (after)	-0.13	0.0068	-19.48

Table 4.16: Summary of coefficients of the fixed-effects predictors for the model without neighborhood density and neighbor frequency. Other aspects of the model are the same as the baseline model. The model contains 12,986 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

Predictor	β	Std. Error	t value
(intercept)	-1.37	0.023	-60.36
baseline duration	0.68	0.050	13.63
bigram probability (before)	-0.014	0.0014	-10.58
bigram probability (after)	-0.023	0.0012	-18.78
disfluency (after): T	0.36	0.0099	36.66
frequency	-0.030	0.0044	-6.90
part of speech: ADV	-0.061	0.066	-0.92
part of speech: N	0.022	0.019	1.13
part of speech: V	-0.097	0.020	-4.81
phonotactic probability (phoneme)	-0.046	0.018	-2.60
speech rate (before)	-0.080	0.0066	-12.12
speech rate (after)	-0.13	0.0068	-19.47

(a) Model without neighborhood density, neighbor frequency and biphone probability

Predictor	β	Std. Error	t value
(intercept)	-1.37	0.023	-60.33
baseline duration	0.68	0.050	13.40
bigram probability (before)	-0.014	0.0014	-10.60
bigram probability (after)	-0.023	0.0012	-18.77
disfluency (after): T	0.36	0.0099	36.64
frequency	-0.031	0.0044	-6.98
part of speech: ADV	-0.057	0.066	-0.88
part of speech: N	0.019	0.019	0.97
part of speech: V	-0.10	0.020	-4.96
phonotactic probability (biphone)	-0.025	0.0089	-2.76
speech rate (before)	-0.080	0.0066	-12.14
speech rate (after)	-0.13	0.0068	-19.48

(b) Model without neighborhood density, neighbor frequency and phoneme probability

Table 4.17: Summary of coefficients of the fixed-effects predictors in two model variants, one without neighborhood metrics and biphone probability and the other without neighborhood metrics and phoneme probability. Other aspects of the models are the same as the baseline model. Both models contain 12,986 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.022	-60.77
neighborhood density	-0.0056	0.0010	-5.39
neighbor frequency	0.0035	0.025	0.14
baseline duration	0.68	0.049	13.72
bigram probability (before)	-0.014	0.0014	-10.54
bigram probability (after)	-0.023	0.0012	-18.79
disfluency (after): T	0.36	0.0099	36.57
frequency	-0.031	0.0043	-7.08
part of speech: ADV	-0.071	0.063	-1.11
part of speech: N	0.013	0.019	0.700
part of speech: V	-0.099	0.020	-5.03
phonotactic probability (phoneme)	0.030	0.022	1.33
speech rate (before)	-0.080	0.0066	-12.14
speech rate (after)	-0.13	0.0068	-19.53

(a) Model without biphone probability

Predictor	β	Std. Error	t value
(intercept)	-1.36	0.022	-60.95
neighborhood density	-0.0053	0.0010	-5.21
neighbor frequency	0.0040	0.025	0.16
baseline duration	0.68	0.050	13.67
bigram probability (before)	-0.014	0.0014	-10.54
bigram probability (after)	-0.023	0.0012	-18.80
disfluency (after): T	0.36	0.0098	36.58
frequency	-0.031	0.0044	-7.06
part of speech: ADV	-0.068	0.064	-1.07
part of speech: N	0.016	0.019	0.82
part of speech: V	-0.097	0.020	-4.95
phonotactic probability (biphone)	0.0094	0.011	0.85
speech rate (before)	-0.080	0.0066	-12.13
speech rate (after)	-0.13	0.0068	-19.53

(b) Model without phoneme probability

Table 4.18: Summary of coefficients of the fixed-effects predictors in two model variants, one without biphone probability and the other without phoneme probability. Other aspects of the models are the same as the baseline model. Both models contain 12,986 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

Predictor	β	**Std. Error**	t **value**
`(intercept)`	-1.36	0.022	-61.00
`neighborhood density`	-0.0048	0.00082	-5.82
`neighbor frequency`	0.0077	0.025	0.31
`baseline duration`	0.67	0.049	13.67
`bigram probability (before)`	-0.014	0.0014	-10.56
`bigram probability (after)`	-0.023	0.0012	-18.79
`disfluency (after): T`	0.36	0.0098	36.58
`frequency`	-0.031	0.0043	-7.19
`part of speech: ADV`	-0.061	0.063	-0.97
`part of speech: N`	0.016	0.019	0.83
`part of speech: V`	-0.097	0.020	-4.98
`speech rate (before)`	-0.080	0.0066	-12.13
`speech rate (after)`	-0.13	0.0068	-19.53

Table 4.19: Summary of coefficients of the fixed-effects predictors for the model without phoneme probability and biphone probability. Other aspects of the model are the same as the baseline model. The model contains 12,986 tokens of 539 word types. "β" denotes the mean estimation of the coefficient; "Std. Error" denotes the standard error in the estimation of the coefficient.

4.5.2 Individual speaker differences

One aspect of the model that has not been discussed yet is the effect of individual speakers. Speaker is always present in the model as a random effect regarding the intercept, meaning that the model will adjust the intercept for each speaker in order to account for by-speaker variance in inherent word duration. It should be noted that in all the models that have been built, the coefficients of the fixed-effects terms are not adjusted for individual speakers. Thus, the model assumes that the same amount of change in neighborhood density should give rise to the same amount of variation (in the same direction) in word duration for all speakers. This is obviously a simplified assumption, though the fact that neighborhood density reaches a significant effect with a common coefficient for all speakers suggests that the effect holds across speakers.

To examine the amount of individual speaker differences regarding the neighborhood effects, I add a random effect of speaker regarding neighborhood density to the baseline model. Thus in addition to adjustment to the intercept, the model also assigns by-speaker adjustment to the coefficient of neighborhood density. In this model, neighborhood density has a coefficient of -0.0056, highly similar to the one in the baseline model (-0.0057), and a t value of -5.10. The standard deviation of by-speaker adjustment for density coefficient is relatively small (0.0011), indicating that the individualized density coefficient is still reliably smaller than zero. Figure 4.12 plots the individualized partial effects of neighborhood density in the model, against scatterplots of word duration and density. Comparison with the baseline model shows that there is no gain in model performance by adding this random term (Chisq=3.02, Chi Df=2, p=0.22).

To conclude, the shortening effect of neighborhood density is robust across the set of speakers. It is not necessarily true that every speaker has this tendency in their production, and it is even less likely that all speakers with this tendency are subject to it to the same degree. However, the pattern must be present in most speakers' production, so that it manifests as a general effect above individual variance; and it must be relatively regular, so that accounting for individualized degrees of the effect is not rewarding.

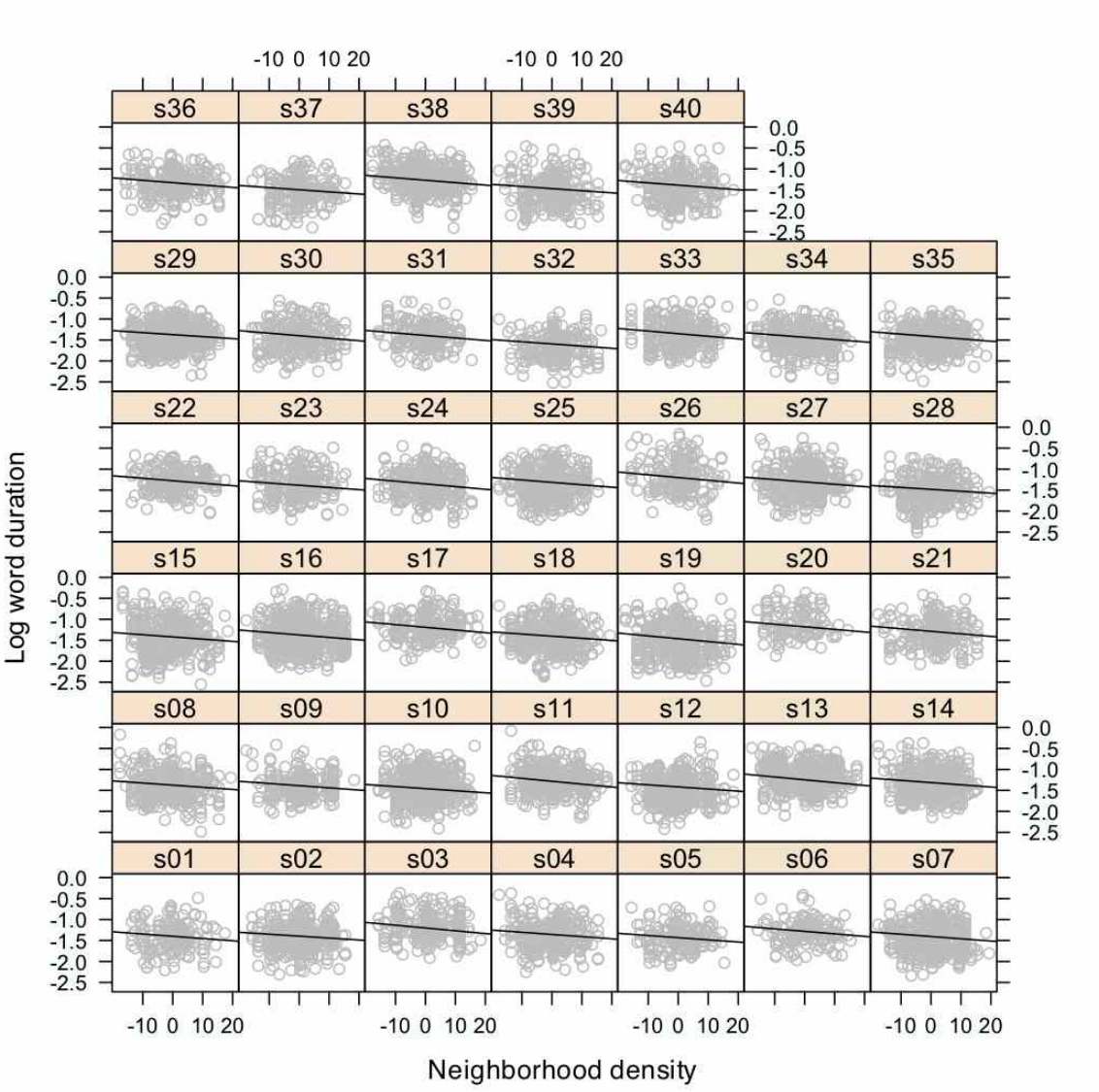


Figure 4.12: Individualized partial effects of neighborhood density on word duration. The grey circles in the background are scatterplots of duration and neighborhood density for each speaker (based on the 12,986-token dataset for the baseline model). Y axis denotes log word token duration; X axis denotes centered neighborhood density from HML. The solid lines represent the individualized partial effects of neighborhood density, based on a model with a random speaker effect regarding the coefficient of neighborhood density. Other aspects of the model are the same as the baseline model.

4.5.3 Theoretical implications

This study starts with two hypotheses, a speaker-oriented hypothesis and a listener-oriented hypothesis. Results from the corpus study show that neighborhood density has a strong effect on word duration. High-density words are realized with shorter durations than low-density words in conversational speech, even though the neighbors may not be mentioned in the context. This finding provides clear evidence for the speaker-oriented hypothesis. Furthermore, the fact that phonotactic probability alone cannot account for the observed shortening effect suggests that the locus of the shortening effect cannot be solely in lower-level articulation. At least part of the effect must be ascribed to higher-level processing (lexical selection, phonological encoding).

The current results provide no support for the listener-oriented hypothesis. In general, words with shorter durations are less intelligible than words with longer durations. In addition, shorter words may also manifest other types of phonetic reduction that are usually linked with lower intelligibility. Thus durational shortening can hardly be associated with promoting intelligibility and therefore current findings do not support the listener-oriented hypothesis. It is possible, though, that speakers are already accommodating listeners by lengthening high-density words but the adjustment is countered by a stronger opposite tendency borne out of the speakers' own production systems. Such a possibility remains to be tested in future research.

As for neighbor frequency, no effect is found in any model that has been tested so far. Consequently, current findings do not support or contest the predictions of either hypothesis regarding neighbor frequency.

To conclude, the corpus study on word duration reveals a strong negative effect of neighborhood density, which provides evidence for a speaker-oriented hypothesis of speech variation, but not for a listener-oriented hypothesis. The finding seems to run counter to earlier results regarding vowel hyperarticulation in high density words (Munson and Solomon 2004; Wright 1997, 2004). But the dependent variable in those studies was degree of vowel dispersion, not duration. Therefore, I conducted a second corpus study, on the effects of neighborhood structure on vowel dispersion. This study will be presented in the next chapter.

Chapter 5

Phonological Neighborhood Density and Vowel Dispersion

In the previous chapter, I have shown that words from dense neighborhoods are produced with shorter durations than those from sparse neighborhoods. In this chapter, I will present the second corpus study, which is on the effects of neighborhood structure on vowel dispersion. The speaker-oriented hypothesis predicts that high-density words will be produced with a more *centralized* vowel space than low-density words, whereas the listener-oriented hypothesis predicts that high-density words should be produced with a more *dispersed* vowel space. These two hypotheses are tested by a mixed-effect model on degree of vowel dispersion.

Using the method proposed in Bradlow et al. (1996), which was also adopted by Munson and Solomon (2004) and Wright (1997), absolute degree of vowel dispersion is measured as the Euclidean distance between the vowel token and the center of vowel space on a F1-F2 plane. To control for vowel-specific inherent degree of dispersion (e.g. [iy] is on average farther away from the center than most other vowels), a K-dispersion measure is developed, which represents the number of standard deviations from the mean degree of dispersion of the particular vowel. A positive K-dispersion value indicates that the token is more dispersed than an average production of the target vowel, and a negative K-dispersion value indicates the token is less dispersed (i.e. more centralized) than the average production.

A mixed-effect model is built on K-dispersion, using a subset of the database presented in Chapter 3 that only includes words with peripheral monophthong vowels (i.e. [aa], [ae], [eh], [ey], [ih], [iy]), [ow], [uh], [uw]) in the citation forms. To preview the results, words from dense neighborhoods are produced with more centralized vowel tokens than words from sparse neighborhoods, which is consistent with the shortening effect of neighborhood density found in Chapter 4. However, there is also a tendency for words with high-frequency neighbors to be produced with more dispersed vowels, but the significance of this effect depends on how neighbor frequency is coded. When CELEX-based neighborhood measures are used, the effect of neighbor frequency fails to reach significance. Overall, the major finding of this study, i.e. the centralizing effect of neighborhood density, supports the speaker-oriented hypothesis over the listener-oriented hypothesis.

The organization of the rest of the chapter is as follows: I will first recapitulate the two hypotheses for the current study (§5.1), and then describe the database and the procedure for coding variables (§5.2); next I will present the results of the baseline model (§5.3) and alternative models (§5.4), and end the chapter with a discussion of the model results (§5.5).

5.1 Predictions

Two sets of hypotheses were formulated in Chapter 2 for the effects of neighborhood structure on pronunciation variation. The speaker-oriented hypothesis predicts that high-density words are more *reduced* than low-density words because they are easier to produce; the listener-oriented hypothesis predicts that high-density words are more *hyperarticulated* than low-density words in order to compensate for their perceptual difficulty. In terms of vowel production, phonetic reduction would correspond to vowel centralization and vowel space contraction, and hyperarticulation corresponds to vowel dispersion and vowel space expansion. Below are the complete arguments of the two hypotheses, regarding the effects of neighborhood density and neighbor frequency on vowel dispersion.

(1) Speaker-oriented hypothesis regarding neighborhood effects on vowel dispersion

   (a) High-density words are easier to produce than low-density words. Words with high-frequency neighbors are easier to produce than those with low-frequency neighbors.

   (b) Easy-to-produce words tend to have more reduced vowels.

   (c) Therefore high-density words and words with high-frequency neighbors will be realized with *more reduced* vowels than low-density words and words with low-frequency neighbors, respectively.

(2) Listener-oriented hypothesis regarding neighborhood effects onvowel dispersion

   (a) High-density words are harder to recognize than low-density words. Words with high-frequency neighbors are harder to recognize than those with low-frequency neighbors.

   (b) Words with more dispersed vowels are more intelligible.

   (c) Speakers monitor for listeners' needs and modify their speech accordingly.

   (d) Therefore high-density words and words with high-frequency neighbors will be realized with *more dispersed* vowels than low-density words and words with low-frequency neighbors, respectively.

Vowel	aa	ae	ah	aw	ay	eh	er	ey	ih	iy	ow	oy	uh	uw
*(IPA)**	a	æ	ə, ʌ**	aʊ	aɪ	ɛ	ɚ	e	ɪ	i	o	oɪ	ʊ	u
Type	56	58	38	10	39	45	24	63	60	48	47	6	15	31
Token	1325	915	837	225	1209	1443	783	1458	1740	918	896	40	1042	411

Table 5.1: Number of word types and tokens per vowel type (according to the citation form) in the duration database (540 word types and 13,242 tokens in total).

**Note that the IPA symbols are only provided as the phonetic reference for the ARPABET symbols, and should not be taken as the exact IPA transcription of the sounds.*
*** Since the citation forms in the Buckeye corpus only make use of one mid-central unrounded vowel, [ah], both /ə/ and /ʌ/ are transcribed as [ah].*

5.2 Data

5.2.1 Database

The database for the previous study on word token duration (i.e. the "duration database"; see the description in §3.1.2 and §4.2.1) contains a total number of 540 CVC monomorphemic content words, represented by 13,242 tokens from 40 speakers. Altogether 14 different vowels appear in the citation forms of words in the duration database (see Table 5.1).

A subset of words (tokens) from the duration database was used in the study on vowel dispersion. Words with diphthongs [aw] (e.g. *couch*), [ay] (e.g. *wide*) and [oy] (e.g. *boil*) and words with mid-central vowels [ah] (e.g. *mud*) and [er] (e.g. *bird*) were not included in the analysis. Diphthongs were excluded because vowel dispersion is hard to measure when the articulators are moving from one vowel position to another[1]. Mid-central vowels were excluded because their overall centrality prevents reliable estimation of vowel dispersion, as the dispersion measure will be highly sensitive to noise in the estimation of center of vowel space.

Thus, only word tokens with non-central monophthong vowels in the citation form were retained for the analysis. There are nine such vowels in the duration database: [aa], [ae], [eh], [ey], [ih], [iy], [ow], [uh], and [uw]. Further data exclusion was based on the availability of reliable formant frequency measures. Altogether 133 word tokens were removed for not having vowels in actual pronunciation, according to the phonetic transcription in the corpus. In addition, the tokens from speaker s35 (n=233) were all excluded, because the transcription

[1]Words with the vowel [ey] (e.g. *fail*) were retained because [ey] is pronounced more like monophthongs in American English. In addition, even in cases when [ey] is diphthongized as /eɪ/, the beginning and end points of the vowel trajectory are relatively close to each other (compared with other diphthongs such as [aw], [ay] and [oy]) and the distance from the center of vowel space can be reliably measured.

of this speaker's speech contains massive errors[2]. Another 74 tokens were excluded because the beginning and end time of the vowel periods could not be reliably located from corpus transcription. Finally, 23 word tokens with vowel periods shorter than the window size (25ms) of the formant-tracking program were removed, as well as 29 tokens for which the formant-tracking program failed to return formant frequencies.

Table 5.2 summarizes all steps of data exclusion that are described above. The final database for the vowel study (i.e. the "vowel database") contains 418 target words, represented by 9,656 tokens from 39 speakers. On average, each speaker has produced 81.26 words (s.d.=18.06) and 247.59 tokens (s.d.=100.18). Table 5.3 shows the number of word types and tokens of each vowel type in the vowel database.

Reason for exclusion	**Number of items removed**	
	Type	***Token***
Type-based:		
Diphthong ([aw], [ay], [oy]) in the citation form	55	1474
Central vowel ([ah], [er]) in the citation form	62	1620
Token-based:		
No vowel in the transcription	– *	133
Speaker s35	– *	233
Vowel periods cannot be located	– *	74
Vowel periods shorter than 25ms	– *	23
No formant frequencies returned by Praat	– *	29
Remainder	418	9656

Table 5.2: Number of word types and tokens removed at each step of data exclusion when compiling the vowel database from the duration database.
**For token-based exclusion, only numbers of eliminated tokens are shown, though the exclusion also caused accidental removal of 5 (i.e. 540 - 55 - 62 - 418) word types.*

5.2.2 Outcome variable

The outcome variable of the current study is K-dispersion, which is a vowel-specific standardized score of degree of dispersion. Following Bradlow et al. (1996), absolute degree of vowel dispersion is calculated as the Euclidean distance from the vowel token to the center of vowel space on the F1-F2 plane. However, this measure is not normally distributed in the vowel database, which violates an underlying assumption for outcome variable in a

[2]Two out of 7 of s35's transcription files were contaminated. However, word-level time labels were not affected, so data from these two files were still included in the database for word duration.

Vowel	aa	ae	eh	ey	ih	iy	ow	uh	uw
(IPA)	a	æ	ɛ	e	ɪ	i	o	ʊ	u
Type	55	58	44	62	60	46	47	15	31
Token	1270	880	1362	1401	1650	869	861	978	385

Table 5.3: Number of word types and tokens per vowel type (according to the citation form) in the vowel database (418 word types and 9,656 tokens in total).

mixed-effect model. In view of this, absolute degree of vowel dispersion is then converted to a standardized distance score (z-score) within each vowel type, i.e. K-dispersion, which is normally distributed in the database. The derivation of K-dispersion is presented in more detail in the rest of this section.

5.2.2.1 Step 1: Identifying the vowel period

The identification of vowel periods in target word tokens was based on word-level phonetic transcription in the Buckeye corpus. As mentioned above, 133 tokens that had no vowels in their phonetic transcription were excluded from the analysis. Most of the remaining tokens contained one and only one vowel in the transcription, and the identification of vowel period was straightforward. There were also a small number of tokens (n=77) that had more than one vowel in the transcription. Vowel periods in these tokens were hand coded by the author of the current work. For example, if a token of the word *death* was transcribed as [d eh ae th] in the corpus, then [eh ae] would be coded as the vowel period; but if the word *run* was transcribed as [er r ah n], only the second vowel, [ah], was recognized as the realization of the target vowel.

The time labels of the beginning and end time of the vowel period were obtained from phone-level transcription in the Buckeye corpus. 74 tokens were removed because the time labels could not be reliably located (see Table 5.2), due to the mismatch between word-level and phone-level transcriptions in the corpus.

5.2.2.2 Step 2: Measuring formants

An automatic Praat (Boersma and Weenink 2008) script was run to measure the average F1 and F2 in the middle 50% of the vowel period. The script called the LPC function in Praat, with parameters set to find 5 formants, with a 25ms window length, 2.5ms step size and 50dB dynamic range. Maximum formant was set to 5,000Hz for male speakers and 5,500Hz for female speakers.

Outlier formant measures, which were 2.5 standard deviations away from the mean values of the particular speaker×vowel combination, were manually checked and corrected, if

		aa	ae	eh	ey	ih	iy	ow	uh	uw
Dur (ms)	**f**	119	168	97	109	84	110	139	76	118
		(14)	(32)	(24)	(17)	(15)	(26)	(32)	(20)	(30)
	m	122	152	89	111	84	107	124	83	121
		(28)	(21)	(17)	(19)	(17)	(27)	(23)	(20)	(19)
F1 (Hz)	**f**	747	801	627	505	505	400	582	510	420
		(67)	(90)	(73)	(51)	(46)	(42)	(61)	(52)	(45)
	m	600	611	510	422	422	330	486	435	355
		(77)	(82)	(54)	(36)	(28)	(28)	(63)	(35)	(28)
F2 (Hz)	**f**	1317	1832	1818	2317	2013	2487	1190	1640	1713
		(128)	(80)	(86)	(126)	(139)	(132)	(81)	(104)	(223)
	m	1107	1644	1550	1949	1717	2097	1049	1365	1456
		(104)	(105)	(112)	(108)	(130)	(130)	(119)	(141)	(204)

Table 5.4: Average vowel durations and formant frequencies, by vowel type by speaker sex, calculated from the vowel database. Vowel names represent pronunciation in the citation forms (which may or may not agree with the phonetic transcription). Mean values are averaged over the means of individual speakers in each sex. Values in parentheses are the standard deviations in by-speaker means.

necessary, by the author of the current work. Table 5.4 shows the average vowel durations and formant frequencies of men and women. As found in a previous study (Yao, Tilsen, Sprouse, and Johnson 2010), vowel tokens in the Buckeye corpus are significantly shorter than those elicited from word-reading tasks. In the current database, average vowel duration is 107ms, almost half as short as the average vowel duration (∼200ms) in single-word production as found by Hillenbrand, Getty, Clark, and Wheeler (1995). Furthermore, if we plot the average vowel space of speakers in the current database (see Figure 5.1), we can see that the vowel space resembles a parallelogram more than a trapozoidal or triangular shape that is traditionally associated with English vowel space (Hillenbrand et al. 1995; Peterson and Barney 1952), and the change in shape is mostly due to the unrounding and/or fronting of back vowels [uw] and [uh]. The parallelogram vowel space is in accordance with findings from the earlier study (Yao et al. 2010) on the Buckeye corpus as well as another study by Hagiwara (1997) on Californian English speech.

In addition, Figure 5.1 also shows that there is some gender difference in vowel production. Female speakers not only have higher F1 and F2 than male speakers in general, but also produce a slightly more expanded vowel space than male speakers, which is consistent

with findings from previous literature (Hagiwara 1997; Hillenbrand et al. 1995; Yao et al. 2010).

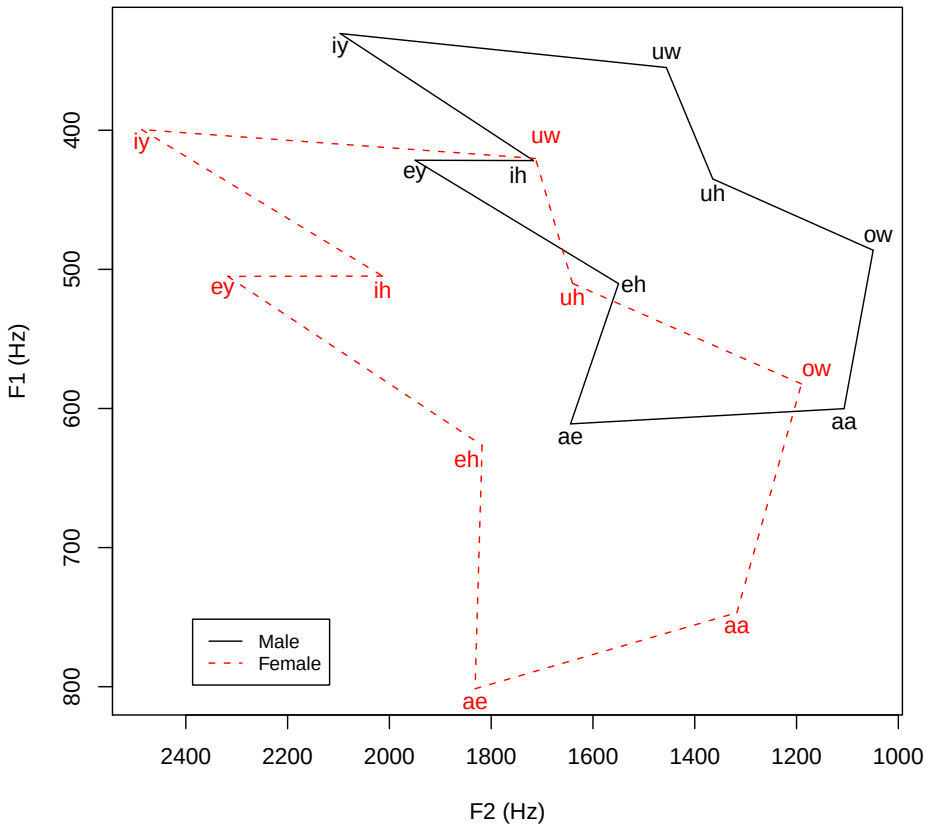


Figure 5.1: Men's (black solid line) and women's (red dotted line) average vowel space, calculated from the vowel database.

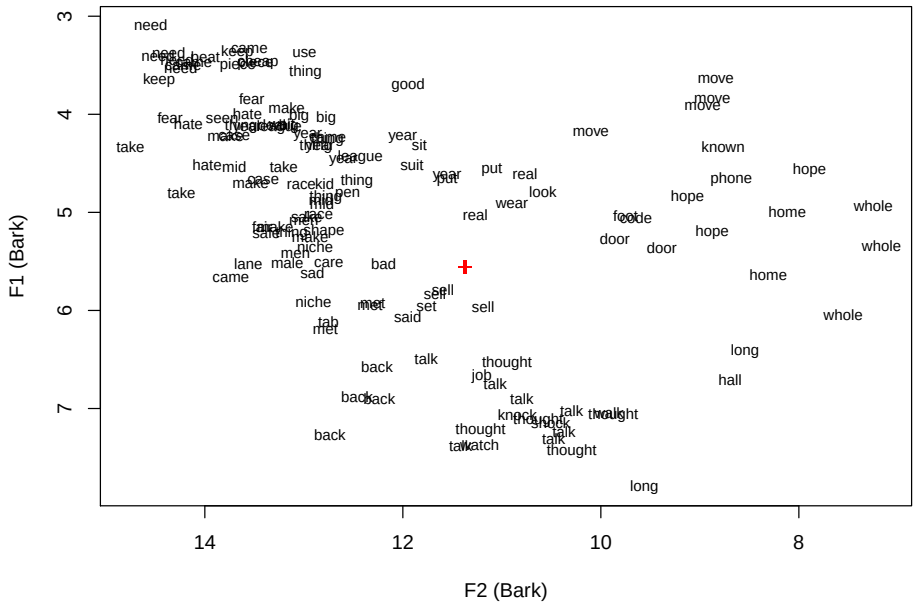
5.2.2.3 Step 3: Calculating absolute degree of dispersion

All frequency measures were converted into the Bark scale (Zwicker and Terhardt 1980), using the formula given in 5.1, where Z is frequency in Bark, and f is frequency in Hz.

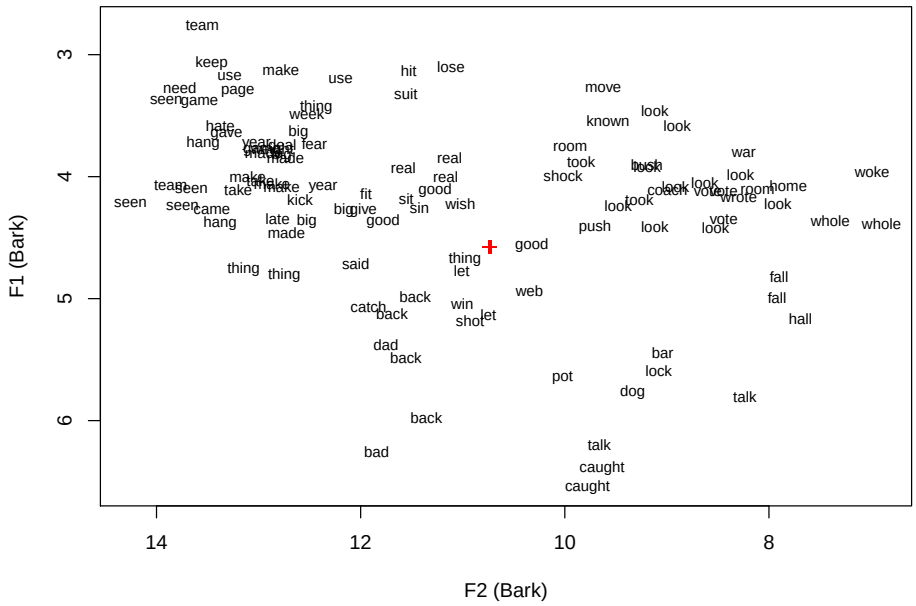
$$Z = \frac{26.81f}{1960 + f} - 0.53 \tag{5.1}$$

Center of vowel space was estimated by the average F1/F2 (in Bark) of four point vowels: [aa], [ae], [iy] and [ow]. Note that different sets of point vowels (or peripheral vowels) have been used in previous studies (e.g. Bradlow et al. (1996) used [aa], [iy] and [ow]; Munson and Solomon (2004) used [aa], [ae], [iy], [ih], [ow] and [uw]). The choice of point vowels largely depends on the shape of the vowel spaces of the particular speakers. In the current study, [aa], [ae], [iy] and [ow] were chosen to represent the four (high/low × front/back) corners of the vowel spaces of Buckeye speakers. In particular, the vowel [ow] was used instead of [uw] as a reference point for high-backness, because [uw] was significantly fronted in Buckeye speakers' speech.

In the current database, the average center of vowel space is (5.90 Bark, 11.62 Bark) for female speakers in the Buckeye corpus and (4.90 Bark, 10.68 Bark) for male speakers. For illustration, Figure 5.1 plots the vowel spaces of two speakers, one female and one male, with each vowel token labeled by the containing word and the center of vowel space marked by a "+" sign.



(a) s05 (female)



(b) s06 (male)

Figure 5.1: Vowel spaces (in Bark) of two speakers, s05 (female) and s06 (male). Formant frequencies of each vowel token are labeled by the containing word. Center of vowel space is marked by a "+" sign.

Next, using the method developed in Bradlow et al. (1996) and widely adopted by previous neighborhood studies (Munson and Solomon 2004; Wright 1997, 2004), absolute degree of vowel dispersion was calculated as the Euclidean distance from the vowel token to the center of vowel space on the F1-F2 plane. The formula is given in 5.2, where $f1$ and $f2$ are the average formant frequencies of the vowel token, and $f1_{center}$ and $f2_{center}$ are the formant frequencies of the center.

$$D = \sqrt{(f1 - f1_{center})^2 + (f2 - f2_{center})^2} \tag{5.2}$$

The range of absolute degree of dispersion in the vowel database is from 0.027 to 8.34, with a median value of 1.86 and a mean value of 1.98 (s.d. = 0.94). However, as shown in Figure 5.2, absolute degree of dispersion is not normally distributed in the dataset. Further analysis shows that a major source of the non-normality is in the vowel-specific inherent degree of dispersion. To illustrate this point, Figure 5.3 shows the by-vowel mean absolute degree of dispersion, separated by speaker sex. As shown in the plot, male speakers and

female speakers have very similar vowel dispersion measures, but there is considerable variation across vowels. On average the vowel [iy] (mean dispersion = 3.28; s.d. = 0.70) is the most distant vowel from the center of vowel space, and [eh] is (mean dispersion = 1.11; s.d. = 0.49) the closest to the center. A t-test confirms that the difference in absolute degree of dispersion between these two vowels is highly significant ($p<0.001$). If we plot the by-vowel distribution of absolute degree of dispersion (Figure 5.4), we can see that the normality is generally improved. In some vowels (e.g. [aa], [ey], [ih]), the distribution is completely normal.

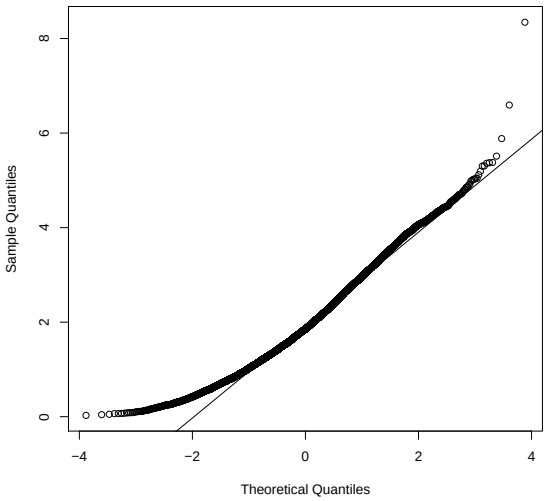


Figure 5.2: Distribution of degree of dispersion in the vowel database (n=9,656).

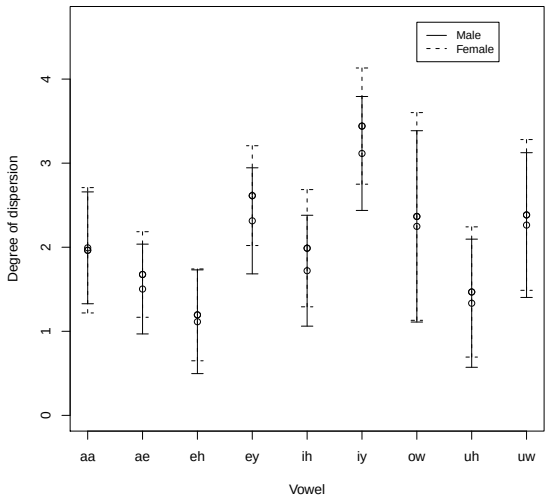
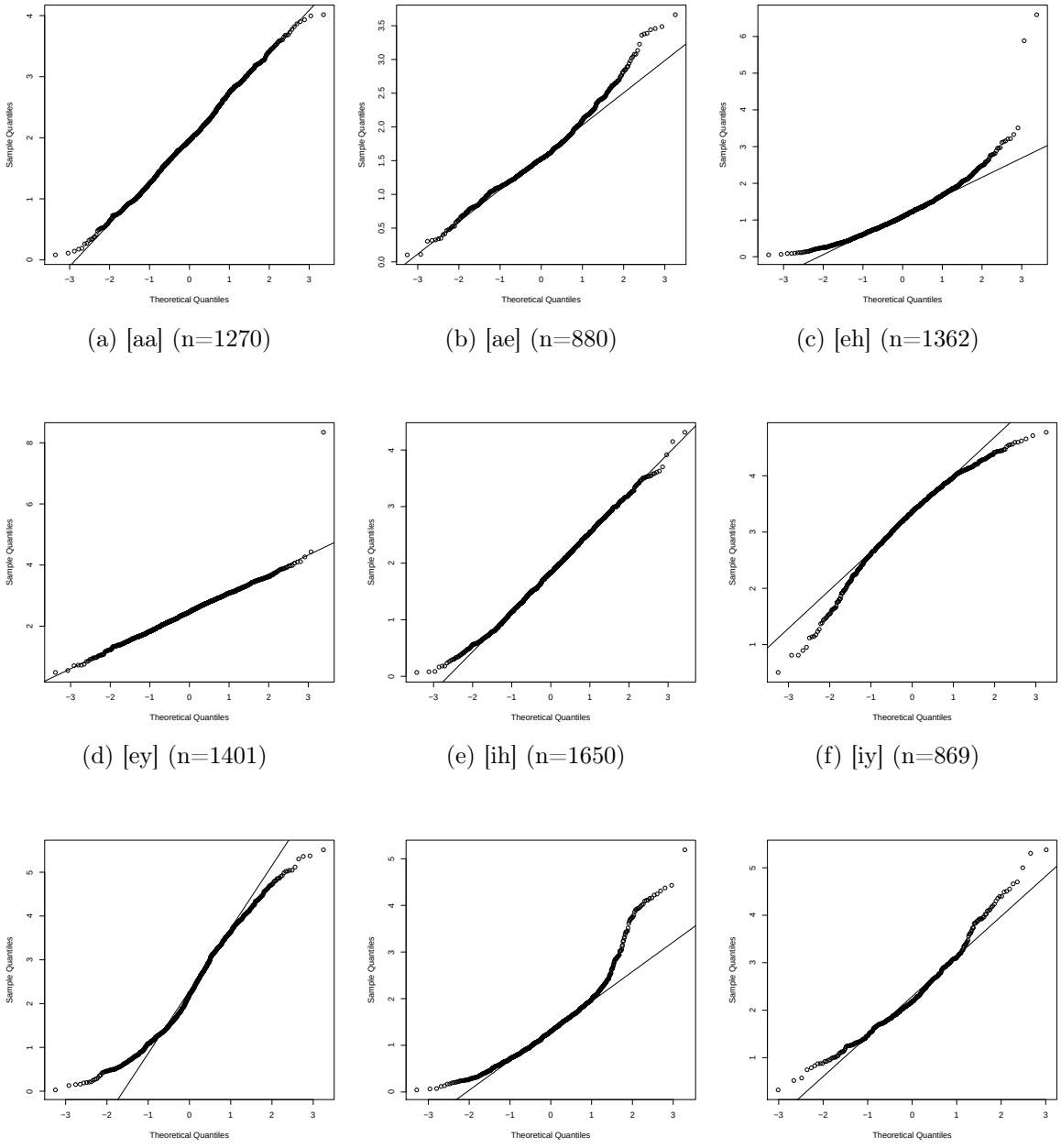


Figure 5.3: Average degree of dispersion by vowel type, separated by speaker sex.



(a) [aa] (n=1270) (b) [ae] (n=880) (c) [eh] (n=1362)

(d) [ey] (n=1401) (e) [ih] (n=1650) (f) [iy] (n=869)

(g) [ow] (n=861) (h) [uh] (n=978) (i) [uw] (n=385)

Figure 5.4: By-vowel distribution of degree of dispersion in the vowel database (n=9,656).

5.2.2.4 Step 4: Converting to K-dispersion

The non-normal distribution of absolute degree of distribution makes it ineligible for being the outcome variable in a mixed-effect model. A solution to this problem is to use a within-vowel standardized dispersion score, i.e. the K-dispersion measure, which normalizes for vowel-specific inherent degree of dispersion. K-dispersion is computed as the number of standard deviations above or below the mean degree of dispersion of the particular vowel type (see the formula given in 5.3). Thus, a positive K-dispersion measure indicates that the vowel token is *farther* from the center of vowel space than an average token of the same vowel type, while a negative K measure indicates the token is *closer* to the center than average.

$$K = (D - D_{mean})/D_{sd}, \tag{5.3}$$

where D is the absolute degree of dispersion, D_{mean} and D_{sd} are the mean and standard deviation of degree of dispersion of the particular vowel type.

The range of K-dispersion in the vowel database is from -3.94 to 9.33, with a median value of -0.05 and a mean value of 3.19×10^{-18} (s.d. = 0.99). Figures 5.5 and 5.6 show the overall distribution and by-vowel mean values of K-dispersion in the database. As shown in the plots, the overall distribution of K-dispersion is highly normal and the average K-dispersion is around zero for all vowel types[3]. Thus, K-dispersion is used as the outcome variable of the vowel dispersion model.

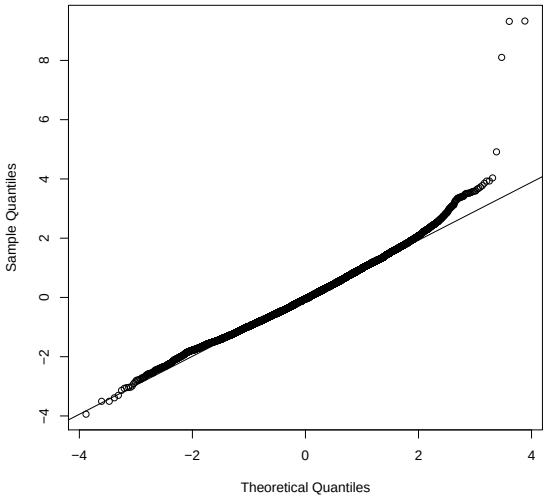


Figure 5.5: Distribution of K-dispersion in the vowel database (n=9,656).

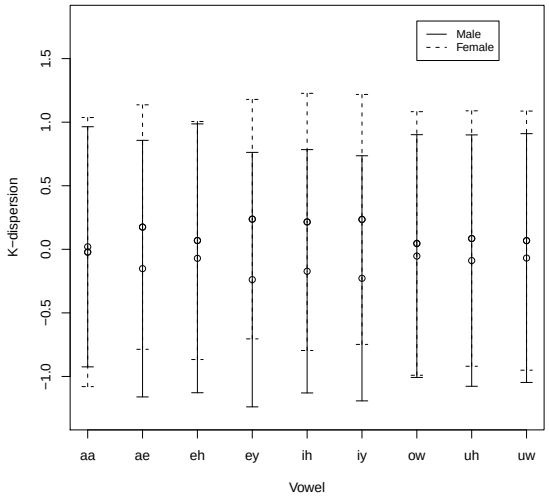


Figure 5.6: Average K-dispersion by vowel type, separated by speaker sex.

[3]However, it should be noted that the transformation does *not* affect by-vowel distributions, because for tokens of the same vowel type (where D_{mean} and D_{sd} are both constant), the transformation is linear.

5.2.3 Predictor variables

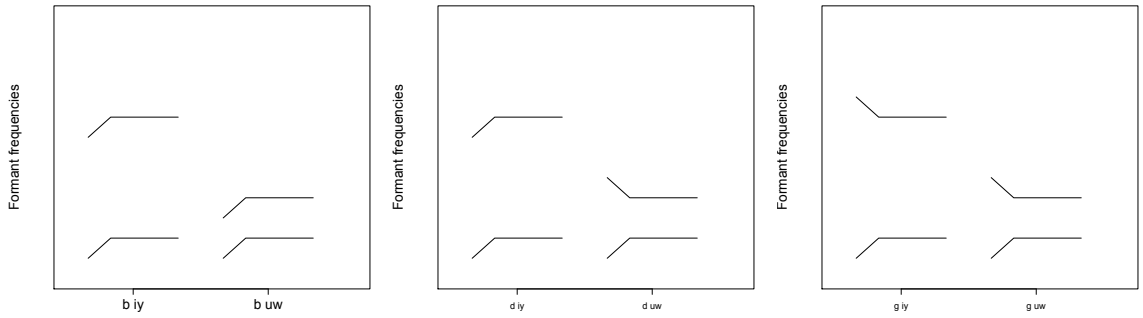
The set of predictor variables in the vowel dispersion model is not completely identical with the set of predictor variables in the word duration model. Changes include the removal of baseline word duration and the addition of vowel type (in the citation form), vowel duration and features of surrounding consonants. In the word duration model, baseline word duration controlled for durational variation due to the phonemic content of the word. In the vowel dispersion model, this type of variation, i.e. by-vowel differences in inherent degrees of dispersion, is largely eliminated by converting absolute degree of dispersion to K-dispersion. But vowel type is still added to the model in case there are any residual by-vowel differences. Vowel duration is added to the model in order to control for duration-dependent vowel undershoot/overshoot (Moon and Lindblom 1994), and it is hypothesized that shorter vowels will tend to be more centralized than longer vowels. Features (place of articulation, manner of articulation, voicing) of the preceding and following consonants are added to control for coarticulation-induced variation. As reviewed in Chapter 2, phonetic context is one of the most important conditioning factors for variation in vowel production. In the following, I will explain how these consonant features might affect vowel production.

A well-established finding in phonetics research is that place of the adjacent consonant affects the patterns of formant transition near the boundary between consonant and vowel (e.g. Delattre, Liberman, and Cooper 1955; Ladefoged 2001). In the current study, even though vowel formants are measured from the middle 50% of the vowel period, the measurement can still be affected by C-V and V-C transitions, because the vowels are relatively short.

As shown in the sketch in Figure 5.7, C-V transition from a labial consonant is featured by rising in both F1 and F2, while the transition from a velar consonant is featured by rising in F1 and falling in F2. Thus, if vowel formants are affected by C-V transitions, we would expect vowels preceded by labial consonants to have lower F2 and those preceded by velar consonants to have higher F2. Following from this, preceding labials should have a centralizing effect on front vowels but a dispersing effect on back vowels, while preceding velars should have a dispersing effect on front vowels but a centralizing effect on back vowels. However, it is not clear whether the degree of such coarticulatory effects would differ between front and back vowels, therefore we have no specific prediction as to the overall effects of preceding labials (or velars) on vowel dispersion.

As for coronal consonants, the transition pattern will depend on the vowel: If the vowel is front (e.g. [iy], [eh] and [ey]), the transition pattern is similar to that of a labial consonant, with both F1 and F2 rising; if the vowel is back (e.g. [aa], [ow], [uw]), the transition is similar to that of a velar consonant, with F1 rising and F2 falling. Thus, we would expect front vowels to be less fronted (i.e. lower F2) and back vowels to be more fronted (i.e. higher F2). In other words, vowels preceded by coronal consonants should be *less* dispersed than vowels preceded by other consonants.

Consonants following vowels have similar effects on vowel production, but in the current study, place of following consonant is predicted to have a smaller effect than place of preceding



(a) Labial (b) Coronal (c) Velar

Figure 5.7: Coarticulatory effect of the place of the preceding consonant on the first two formants of the vowel.

consonant, because final consonants are more often omitted in conversational speech.

Manner of the surrounding consonants may also affect vowel production. Specifically, I considered three manners of articulation: obstruent, nasal and approximant. The greatest effects are probably from approximants, especially from syllable-final [l]. The English language includes four approximants, [l], [r], [y], and [w], among which only two ([l] and [r]) have appeared in syllable-final positions of the target words in the current database. Syllable-final [l] is often velarized in American English, which will result in a lower F2 in the vowel. Thus, we would expect final [l] to be associated with centralization in front vowels but dispersion in back vowels. Furthermore, syllable-final [l] is also a blocking environment for the fronting of back upgliding vowels [uw] and [ow] (Labov, Ash, and Boberg 2006), therefore the lowering in F2 may be more prominent in back vowels than in front vowels, causing the overall effect to be dispersion rather than centralization.

On the other hand, nasal consonants may affect vowel production, too, through nasalization. Nasalized vowels are featured by an extra spectral peak in the lower frequency range, typically lower than the first formant, accompanied by an reduction in the amplitude of the first formant. Consequently, the formant tracking program might recognize the lower spectral peak as the first formant of a nasalized vowel, which will result in a lower F1 measurement. However, since tokens with outlier formant measures were hand-checked and corrected, it is predicted that the effect of surrounding nasal consonants will be small.

Last but not least, vowel production may also be influenced by the voicing feature of the surrounding consonants. A rule of English pronunciation is that vowels preceding voiceless consonants are shorter than those preceding voiced consonants (Klatt 1979). Since duration and vowel dispersion are closely related, voicing of the consonant may affect vowel dispersion, too. However, the effect will probably be largely reduced, if vowel duration is

already controlled for.

Overall it is predicted that (a) vowels surrounded by coronal consonants will tend to be *less* dispersed than those surrounded by labial consonants and velar consonants; (b) vowels followed by approximants will be more dispersed than those followed by obstruents and nasals; (c) vowels followed by voiceless consonants will be shorter than those followed by voiced consonants.

In the following, I will provide a list of descriptions for all predictor variables that are coded for the current dispersion model. These variables include neighborhood density, neighbor frequency, bigram probability, disfluency, familiarity, frequency, features of surrounding consonants (place of articulation, manner of articulation and voicing), orthographic length, part of speech, phonotactic probability, previous mention, speech rate, speaker age, speaker sex, vowel duration and vowel type. For variables that are also existent in the word duration model (i.e. all variables except for features of the surrounding consonants, vowel duration and vowel type), the description is repeated from §4.2.2.

Neighborhood density and neighbor frequency are the critical predictor variables in the current model. Neighborhood density is the number of phonological neighbors under the one-phoneme difference rule. Neighbor frequency is the mean log frequency of neighbors. Both measures come from the Hoosier mental lexicon (Nusbaum et al. 1984), which is in turn based on pronunciations from the Webster pocket dictionary (which has about 20,000 lexical entries) and word frequency from Kučera and Francis (1967) (which is based on the 1-million word Brown corpus).

Bigram probability is calculated as the conditional probability of the current word given the adjacent word, as an estimate for contextual predictability. Each word token is coded with two bigram probability measures: one with the preceding word (i.e. $C(W_{n-1}W_n)/C(W_{n-1})$) and the other with the following word (i.e. $C(W_nW_{n+1})/C(W_{n+1})$). Frequency counts of the neighboring words and the word pairs are based on the entire corpus. Both probability measures are log transformed.

Importantly, only adjacent words in the same stretch are considered as valid context. Thus words at the beginning or end of a stretch will have no bigram probability. The end of a stretch is encountered when one of the following items appears in the transcription: silence longer than half a second, cutoff words (i.e. transcribed with <CUTOFF> labels), words with lexical or phonological errors (<ERROR>), interviewer's speech (<IVER>), and non-linguistic sounds such as laughter (<LAUGH>), non-speech vocal noise (<VOCNOISE>) and environmental noise (<NOISE>).

Disfluency is coded by two logical variables, one for the preceding context and the other for the following context. If the target token is immediately preceded or followed by cutoff words (i.e. transcribed with the <CUTOFF> labels), word errors (i.e. transcribed with the <ERROR> labels) or filled pauses (*um, uh* or *you know*), the corresponding disfluency variable will be coded as TRUE, otherwise it will be FALSE.

Familiarity is the subjective familiarity rating of words. The current metric is from the Hoosier mental lexicon (Nusbaum et al. 1984), which encodes word familiarity from 1 (unknown) to 7 (highly familiar).

Features of the surrounding consonants are coded by six predictor variables, which represent the place of articulation (LABIAL, CORONAL, BACK), manner of articulation (OBS=obstruent, NASAL, APPROX=approximant) and voicing (VOICELESS, VOICED) of the preceding and following consonants. The coding scheme for each consonant is shown in Table 5.5. (Note a distinction is made between syllable-initial "clear L" and syllable-final "dark L", which have different place of articulation.)

Frequency is the usage frequency of the word form. The current measure is from the CELEX frequency dictionary for wordforms, which is in turn based on the 17.9 million-word Collins Birmingham University International Language Database (COBUILD corpus; Sinclair 1987). Log frequency (i.e. $ln(Freq + 1)$) is used in the model.

Orthographic length is the length of the word in letters.

Part of speech is the syntactic category of the target word. Since the Buckeye corpus has no syntactic transcription, I used part of speech tags from the lemma-based CELEX syntax dictionary and all tokens of the same word are coded with the same part of speech. For words with more than entries in the syntax dictionary, usage frequency for each possible part of speech is summed and the most frequent part of speech is recorded for the word (e.g. *time* is coded as a noun instead of a verb). Part of speech of irregular forms were manually added (e.g. *came* is coded as a verb; all past participles are coded as verbs). Words in the final database all belong to one of the four syntactic categories: adjective (A), adverb (ADV), noun (N) and verb (V).

Phonotactic probability refers to the probability of having the exact phonemic composition of a word. In this work, each word is coded with two probability measures: average phoneme probability and average biphone probability. Take the word *cat* for example. The average phoneme probability of *cat* is the mean value of the probability of having [k] in the initial position, the probability of having [ae] in the second position and the probability of having [t] in the third position in the English language. Likewise, average biphone probability of *cat* equals the mean value of the probability of having [k ae] in the first and second positions and the probability of having [ae t] in the second and third positions. All raw probability measures are obtained from the web-based Phonotactic Probability Calculator (Vitevitch and Luce 2004). Log probabilities are used in the model.

Previous mention is coded as a logical variable. If the same word has been produced by the speaker at least once from the beginning of the interview to the point right before the target token, it will be coded as TRUE, otherwise it will be FALSE.

Consonant	*(IPA)*	**Place**	**Manner**	**Voicing**
b	b	LABIAL	OBS	VOICED
ch	ʧ	CORONAL	OBS	VOICELESS
d	d	CORONAL	OBS	VOICED
f	f	LABIAL	OBS	VOICELESS
g	g	BACK	OBS	VOICED
hh	h	BACK	OBS	VOICELESS
jh	ʤ	CORONAL	OBS	VOICED
k	k	BACK	OBS	VOICELESS
l (syllable-initial)	l	CORONAL	APPROX	VOICED
l (syllable-final)	ɬ	BACK	APPROX	VOICED
m	m	LABIAL	NASAL	VOICED
n	n	CORONAL	NASAL	VOICED
ng	ŋ	BACK	NASAL	VOICED
p	p	LABIAL	OBS	VOICELESS
r	ɹ	CORONAL	APPROX	VOICED
s	s	CORONAL	OBS	VOICELESS
sh	ʃ	CORONAL	OBS	VOICELESS
t	t	CORONAL	OBS	VOICELESS
th	θ	CORONAL	OBS	VOICELESS
v	v	LABIAL	OBS	VOICED
w	w	LABIAL	APPROX	VOICED
y	j	CORONAL	APPROX	VOICED
z	z	CORONAL	OBS	VOICED

Table 5.5: Coding consonant features. Only consonants that appear in the citation forms of the target words (n=418) in the current dataset are shown in the table. Place of articulation has three levels: OBS (obstruent), NASAL and APPROX (approximant). Manner of articulation has three levels: LABIAL (bilabial, labial-dental, and labial-velar), CORONAL (dental, alveolar and post-alveolar) and BACK (velar and glottal). Voicing has two levels: VOICED and VOICELESS.

Speaker age is coded as a binary variable in accordance with the age stratification in the corpus. Exactly half the speakers (n=20) were under forty years old (Y) and the other half were above (O) at the time of recording. The actual age of the speakers ranged from late teens to late seventies.

Speaker sex is coded as a binary variable. Half the speakers are male (M) and the other half are female (F).

Speech rate is coded by two numerical variables, one for the preceding part of the local stretch and the other for the following part. To avoid autocorrelation, the target token is *not* included in the calculation of either speech rate measure. Raw speech rate is measured in number of syllables per second. Log transformed speech rate is used in the model.

Vowel duration is the duration of the vowel period in the actual pronunciation (see §5.2.2 for the identification of the vowel period). Log transformed vowel duration is used in the model.

Vowel type is the vowel name in the citation form. As described in §5.2.2, the current word set (n=418) features nine vowels in the citation forms: [aa], [ae], [eh], [ey], [ih], [iy], [ow], [uh] and [ow]. Individual word type and token counts of each vowel type in the dataset are shown in Table 5.3 in §5.2.2.

5.2.4 Summary statistics of all variables

The vowel dispersion model consists of one numerical outcome variable (K-dispersion), two random variables (word and speaker) and 25 fixed-effects predictor variables, among which 12 are numerical and 13 are categorical. Table 5.7 below shows the summary statistics of K-dispersion and all of the numerical predictors in the 9,656-token vowel database. Table 5.6 summarizes for the categorical fixed-effects predictors in the database.

Variable	Min	Max	Median	Mean	s.d.	Log
`K-dispersion`	-3.94	9.33	-0.05	3.19×10^{-18}	-0.99	
`neighborhood density`	3	40	21	21.12	6.95	
`neighbor frequency` [4]	1.28	2.95	2.11	2.10	0.24	
`bigram probability (before)`	0.000079	0.75	0.005	0.026	0.069	Yes
`bigram probability (after)`	0.000079	0.83	0.0047	0.03	0.078	Yes
`familiarity`	2.4167	7	7	6.96	0.13	

continued on next page...

[4] Average neighbor frequency in HML is already log-transformed

Table 5.6 – continued from previous page

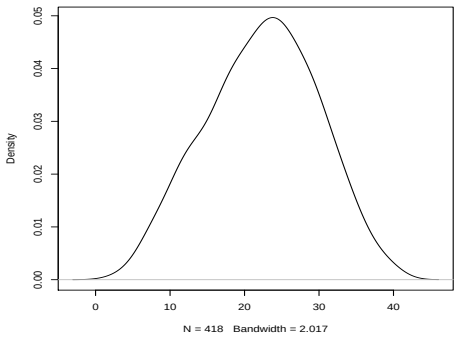
Variable	Min	Max	Median	Mean	s.d.	Log
frequency	0 (per 17.9m)	49655 (per 17.9m)	6750	9534.63	11114.03	Yes
orthographic length	3 (letter)	7 (letter)	4	4.01	0.72	
phonotactic probability (phoneme)	0.012	0.098	0.049	0.049	0.017	Yes
phonotactic probability (biphone)	0.00055	0.016	0.0024	0.0032	0.0025	Yes
speech rate (before)	0.95 (syll/s)	33.33 (syll/s)	5.90	6.20	2.28	Yes
speech rate (after)	0.88 (syll/s)	41.04 (syll/s)	5.19	5.27	1.70	Yes
vowel duration	0.025 (s)	0.49 (s)	0.095	0.11	0.056	Yes

Table 5.6: Summary statistics for numerical predictor variables in the vowel database, based on 9,656 tokens.

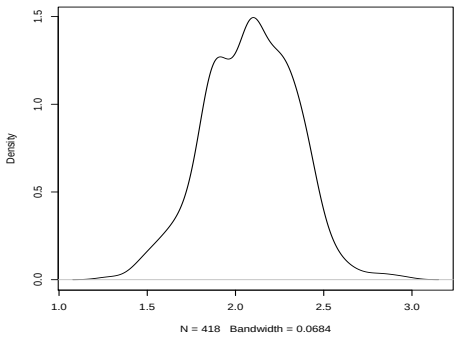
Variable name	Token counts at each level
disfluency (before)	F = 9544; T = 112
disfluency (after)	F = 9189; T = 467
manner of consonant (before)	APPROX = 1736; NASAL = 1146; OBS = 6774
manner of consonant (after)	APPROX = 1747; NASAL = 1492; OBS = 6417
part of speech	A = 2152; ADV = 507; N = 2849; V = 4148
place of consonant (before)	BACK = 2065; CORONAL = 4697 ; LABIAL = 2894
place of consonant (after)	BACK = 3564; CORONAL = 4866; LABIAL = 1226
previous mention	F = 2808; T = 6848
speaker age	O = 5267; Y = 4389
speaker sex	F = 4720; M = 4936
voicing of consonant (before)	VOICED = 4912; VOICELESS = 4744
voicing of consonant (after)	VOICED = 5936; VOICELESS = 3720
vowel type	[aa] = 1270; [ae] = 880; [eh] = 1362; [ey] = 1401; [ih] = 1650; [iy] = 869; [ow] = 861; [uh] = 978; [uw] = 385

Table 5.7: Summary statistics for categorical predictor variables in the vowel database, based on 9,656 tokens.

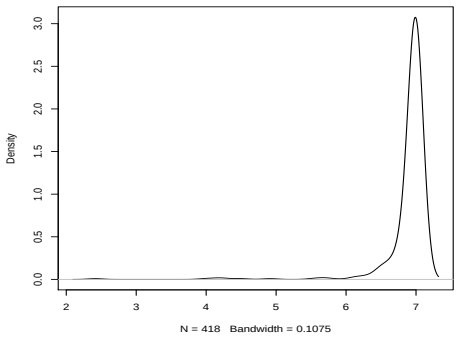
The next set of plots (Figure 5.8 and 5.9) shows the distribution of numerical predictor variables in the database. Because some variables are properties of the word (e.g. frequency) while others are properties of the token (e.g. vowel duration), a distinction is made between type-based and token-based variables. For type-based variables, distribution over the set of words (n=418) in the database is shown, while for token-based variables, distribution over the set of tokens (n=9,656) is shown. If a variable is log transformed, distributions of both the raw values and the log values are shown. As can be seen in the plots, most durational and probabilistic variables achieve more normal distributions after log transformation.



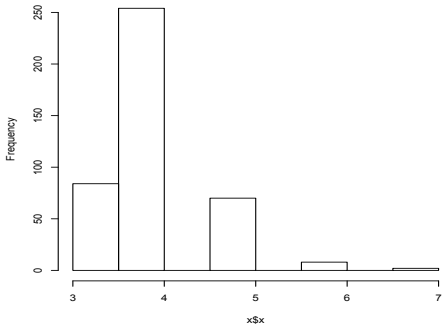
(a) Neighborhood density



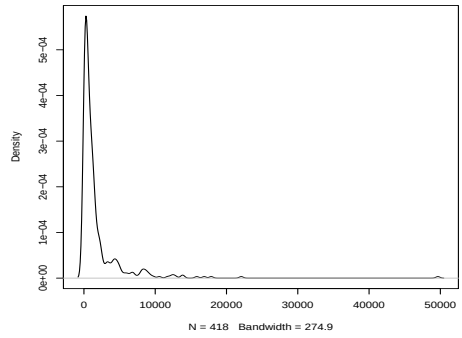
(b) Neighbor frequency



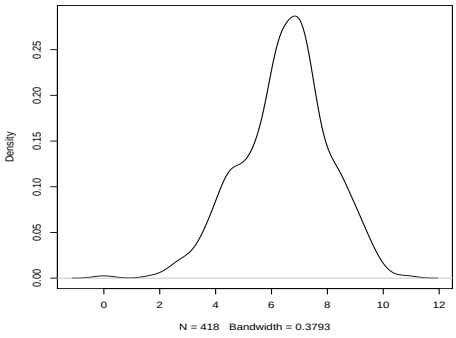
(c) Familiarity



(d) Orthographic length

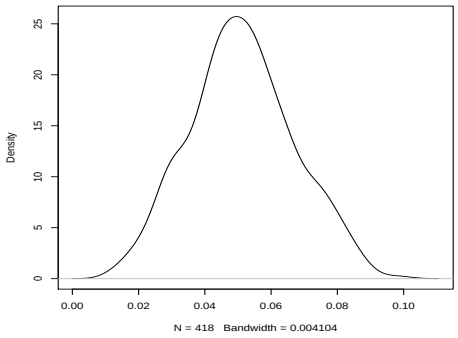


(e) Raw frequency

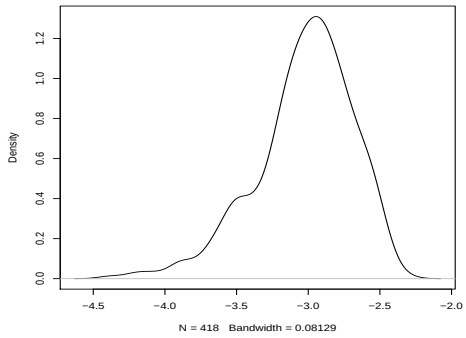


(f) Log frequency

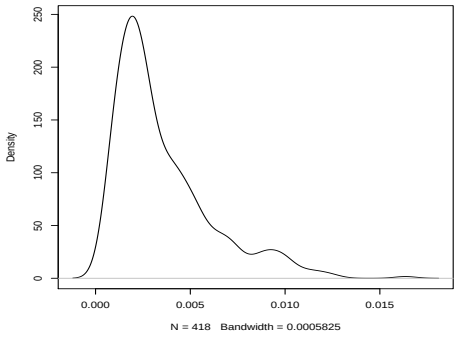
Figure 5.8: Distribution of type-based variables in the vowel dispersion model. Distributions are based on word types (n=418). For orthographic length, histogram is shown. For all other variables, probability functions are shown. (Continued on the next page.)



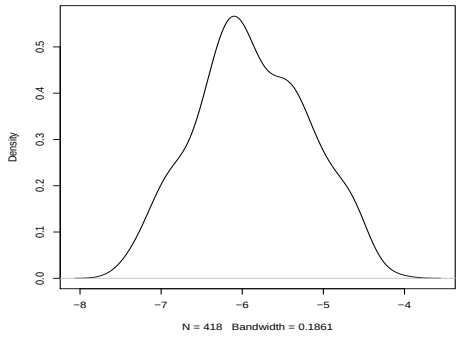
(g) Raw phonotactic probability (phoneme)



(h) Log phonotactic probability (phoneme)

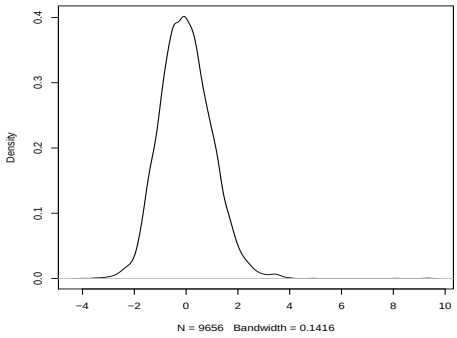


(i) Raw phonotactic probability (biphone)

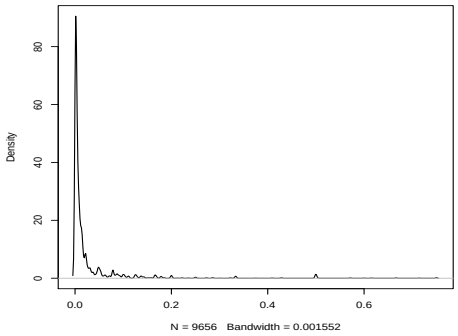


(j) Log phonotactic probability (biphone)

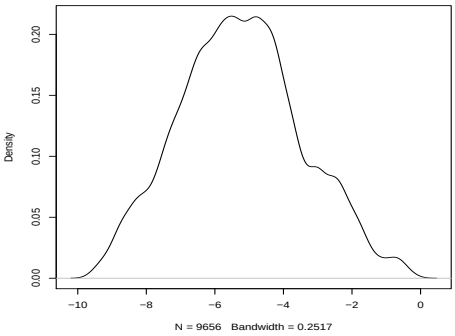
Figure 5.8: Continued: Distribution of type-based variables in the vowel dispersion model. Distributions are based on word types (n=418). For orthographic length, histogram is shown. For all other variables, probability functions are shown.



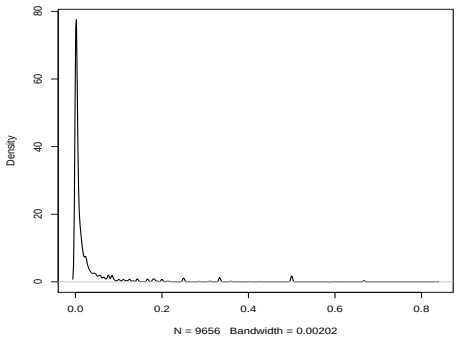
(a) K-dispersion



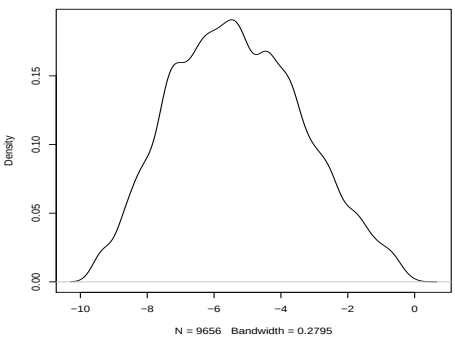
(b) Raw bigram probability (before)



(c) Log bigram probability (before)

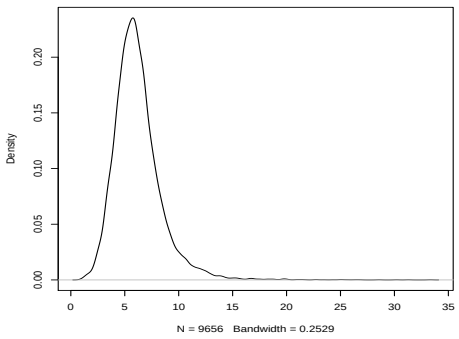


(d) Raw bigram probability (after)

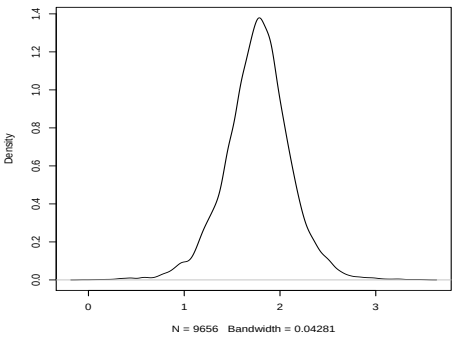


(e) Log bigram probability (after)

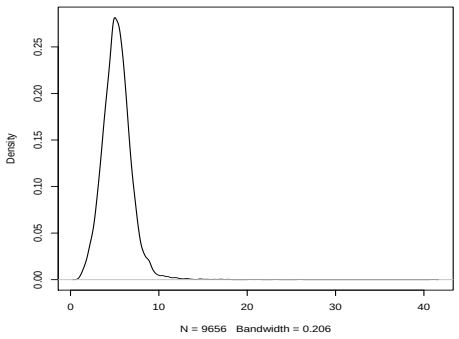
Figure 5.9: Distribution of token-based variables in the vowel dispersion model. For all variables, probability density functions over the token set (n=9,656) are shown. (Continued on the next page.)



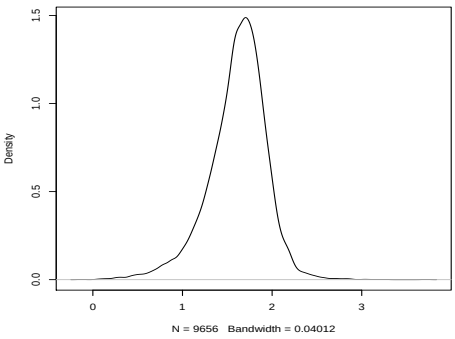
(f) Raw speech rate (before)



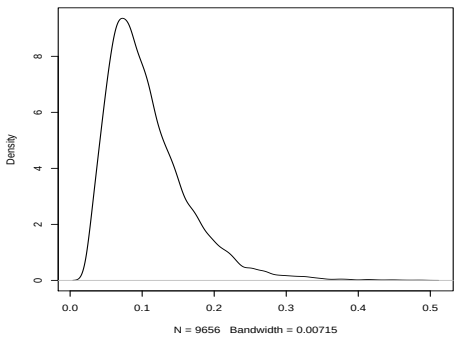
(g) Log speech rate (before)



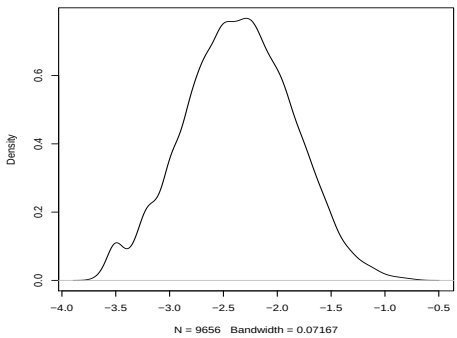
(h) Raw speech rate (after)



(i) Log speech rate (after)



(j) Raw vowel duration



(k) Log vowel duration

Figure 5.9: Continued: Distribution of token-based variables in the vowel dispersion model. For all variables, probability density functions over the token set (n=9,656) are shown.

The following set of tables show the correlations among the fixed-effects predictors in the model. Again, type-based variables are separated from token-based variables, with type-based variables correlated based on the word set (n=418) and token-based variables the token set (n=9,656). Table 5.8 shows all pair-wise correlations among the type-based numerical predictors. As shown in the table, there is considerable correlation between neighborhood density and phonotactic probability (r=0.60 with phoneme probability; r=0.53 with biphone probability), as well as between the two phonotactic variables (r=0.71). This suggests that phonotactic probability needs to be checked as a potential confounding factor for any effects of neighborhood density. Neighbor frequency, on the other hand, has little correlation with neighborhood density, phonotactic probability, or any other numerical predictors (r<0.2 in all cases).

	ND	NFrq	Fam	Frq	OLen	PP_P	PP_Bi
ND	1	0.13	-0.0088	0.05	-0.25	0.60	0.53
NFrq	–	1	-0.007	0.14	0.039	0.16	0.16
Fam	–	–	1	0.25	-0.035	-0.014	0.011
Frq	–	–	–	1	0.035	-0.0069	-0.047
OLen	–	–	–	–	1	-0.37	-0.36
PP_P	–	–	–	–	–	1	0.71
PP_Bi	–	–	–	–	–	–	1

Table 5.8: Pair-wise correlations among type-based numeric variables over the word set (n=418) in the vowel database (ND = neighborhood density; NFrq = neighbor frequency; Fam = familiarity; Frq = (log) frequency; OLen = orthographic length; PP_P = (log) phonotactic probability (phoneme); PP_Bi = (log) phonotactic probability (biphone)).

Table 5.9 shows the ranges and mean values of neighborhood density and neighbor frequency corresponding to different features of surrounding consonants. As shown in the table, all consonant features correspond to similar ranges of neighborhood characteristics, with the mean neighborhood density around 22 and the mean neighbor frequency around 2.1. The only exceptions may be with following approximants (`manner of consonant (after)` = APPROX), which seem to occur in words with higher density (mean = 26.61, range = [8, 40]) than average, and with following labial consonants (`place of consonant (after)` = LABIAL), which seem to occur in words with lower density (mean = 18.75, range = [7, 30]) and lower neighbor frequency (mean = 1.88, range = [1,28, 2.38]) than average. Overall, features of surrounding consonants do not covary with neighborhood characteristics.

Table 5.10 shows all pair-wise correlations among the token-based variables. As shown in the table, there is little correlation among any token-based predictor variables in the model (r<0.2 in all cases).

Consonant feature		ND		NFrq	
		Mean	Range	Mean	Range
manner of consonant (before)	APPROX	23.19	[3, 39]	2.05	[1.48, 2.67]
	NASAL	22.45	[7, 35]	2.09	[1.50, 2.41]
	OBS	22.07	[5, 40]	2.09	[1.28, 2.95]
manner of consonant (after)	APPROX	26.61	[8, 40]	2.21	[1.55, 2.72]
	NASAL	22.27	[8, 38]	2.11	[1.80, 2.81]
	OBS	21.06	[3, 40]	2.04	[1.28, 2.95]
place of consonant (before)	BACK	23.54	[7, 36]	2.10	[1.54, 2.95]
	CORONAL	21.66	[7, 39]	2.03	[1.28, 2.85]
	LABIAL	22.77	[3, 40]	2.14	[1.48, 2.81]
place of consonant (after)	BACK	23.62	[3, 40]	2.04	[1.43, 2.45]
	CORONAL	22.84	[5, 40]	2.17	[1.55, 2.95]
	LABIAL	18.75	[7, 30]	1.88	[1.28, 2.38]
voicing of consonant (before)	VOICED	21.97	[3, 40]	2.05	[1.28, 2.81]
	VOICELESS	22.79	[5, 40]	2.12	[1.43, 2.95]
voicing of consonant (after)	VOICED	22.88	[7, 40]	2.12	[1.28, 2.95]
	VOICELESS	21.72	[3, 40]	2.05	[1.48, 2.85]

Table 5.9: Summary statistics for neighborhood variables in the vowel database, grouped by features of surrounding consonants. ND = neighborhood density; NFrq = neighbor frequency.

	K-Disp	Bigram_before	Bigram_after	Rate_before	Rate_after	VDur
K-Disp	1	-0.028	0.0055	-0.020	-0.055	0.048
Bigram_before	–	1	0.084	-0.015	0.021	-0.032
Bigram_after	–	–	1	0.03	-0.064	-0.24
Rate_before	–	–	–	1	0.16	-0.14
Rate_after	–	–	–	–	1	-0.15
VDur	–	–	–	–	–	1

Table 5.10: Pair-wise correlations among token-based numeric variables over the token set (n=9,656) in the vowel database (K-Disp = K-dispersion; Bigram_before = (log) bigram probability (before) ; Bigram_after = (log) bigram probability (after); Rate_before = (log) speech rate (before); Rate_after = (log) speech rate (after); VDur = (log) vowel duration).

5.3 Model

5.3.1 Model construction

All mixed-effect models in the current study are fitted with the `lmer()` function in the **lme4** package (Bates and Maechler 2010). The construction of the vowel dispersion model followed the procedure as described in §3.2.2 of Chapter 3 on general methodology. An initial model was built on K-dispersion, with all coded predictor variables as fixed-effects terms and speaker and word as two random terms, using the complete 9,656-token vowel database. All numerical variables were centered (i.e. subtracted by the mean) before entering the initial model. Next, the set of fixed-effects predictors (excluding neighborhood characteristics and phonotactic probability) was trimmed in two steps. In the first round of trimming, predictors with absolute t values smaller than 1 were eliminated. Altogether 9 predictors (bigram probability (after), disfluency (before), familiarity, frequency, manner of consonant (before), previous mention, speech rate (before), voicing (before), voicing (after)) were eliminated, and the model was refitted with the remaining predictors. In the second round, predictors with absolute t values smaller than 2 were individually tested for their effects on model fit. If removing a predictor variable failed to affect model fit (i.e. $p>0.05$ in an ANOVA test of model fit), the predictor was removed from the model. In this round, 5 predictors (disfluency (after), orthographic length, part of speech, place of consonant (after), speaker age) were checked and all were eliminated.

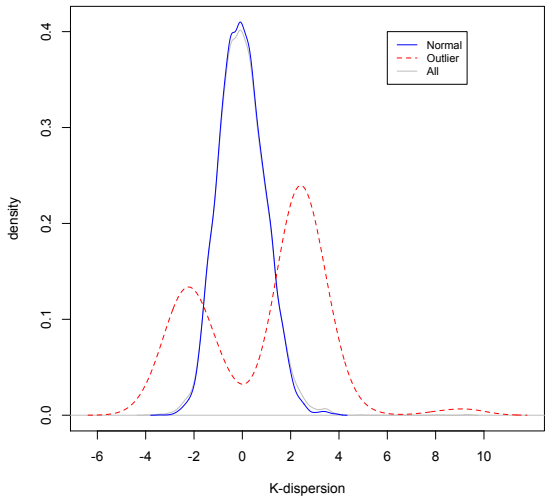


Figure 5.10: Distribution of K-dispersion in the "normal" population and the "outlier" population. Outliers are identified as data points associated with residuals greater than 2.5 standard deviations in the trimmed model. Altogether there are 179 (out of 9,656) outliers.

After the trimming steps, the model was left with 11 predictor variables. The last step

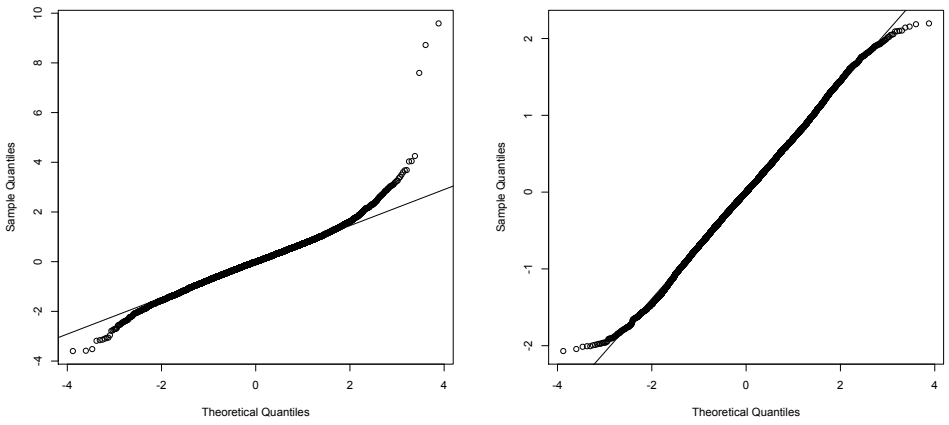
of model construction concerns data points with extremely large residuals, i.e. tokens for which the model has trouble making reasonable predictions. The mean of model residuals is around zero (which is expected for mixed-effect models) and the standard deviation is 0.80, but the range of residuals is [-3.60, 9.59], which is quite large. Tokens with large residuals probably have extreme dispersion measures that are beyond the predicting power of the model. To ascertain if this is the case, I separated tokens with model residuals that were more than 2.5 standard deviations away from zero (i.e. "outlier population", n=179) from the rest of the data points (i.e. "normal population", n = 9,477), and plotted the distribution of K-dispersion in each population against the distribution in the whole population (see Figure 5.10). As shown in Figure 5.10, K-dispersion in the outlier population is centered around two peaks, one on each side of the range [-2, 2], but the distribution in the normal population (which is largely overlapped with the distribution in the complete population) is centered around zero. Recall that K-dispersion is a standardized score, which means that any values outside of [-2, 2] are more than 2 standard deviations from the sample means. Therefore the distribution of K-dispersion in the outlier population suggests that the outlier population is indeed composed of tokens with extreme dispersion values, which are either extremely close to (negative K) or distant from (positive K) the center of vowel space.

To examine the effects of outliers on the model, I refitted the model, excluding the 179 outliers. In the updated model, the range of residuals is much smaller ([-2.07, 2.20]), and the distribution is much more normal (Figure 5.11), but all fixed effects remain unchagned (Table 5.11). In view of this, I will use the updated model, which contains 9,477 word tokens and a trimmed model structure, as the "baseline" model of the current study[5]. Model summary and model evaluation presented in the following sections (§5.3.2 and §5.3.3) are all based on the baseline model.

[5]In this chapter, without further specification, the term "baseline model" refers to the baseline model for **vowel dispersion**.

Predictor	Before removing outliers			After removing outliers		
	β	S. E.	t value	β	S. E.	t value
(intercept)	0.88	0.16	5.60	0.91	0.16	5.78
neighborhood density	-0.018	0.0073	-2.45	-0.018	0.0072	-2.45
neighbor frequency	0.51	0.17	3.06	0.52	0.16	3.13
bigram probability (before)	-0.019	0.0056	-3.35	-0.020	0.0050	-3.96
manner of consonant (after): NASAL	-0.34	0.13	-2.62	-0.35	0.13	-2.77
manner of consonant (after): OBS	-0.43	0.100	-4.18	-0.45	0.100	-4.43
phonotactic probability (phoneme)	0.031	0.17	0.18	0.0051	0.17	0.029
phonotactic probability (biphone)	0.017	0.098	0.17	0.020	0.097	0.21
place of consonant (before): CORONAL	-0.43	0.11	-4.04	-0.45	0.11	-4.18
place of consonant (before): LABIAL	-0.24	0.11	-2.29	-0.24	0.11	-2.25
speaker sex: M	-0.25	0.089	-2.85	-0.27	0.088	-3.07
speech rate (after)	-0.12	0.028	-4.26	-0.11	0.025	-4.25
vowel duration	0.17	0.023	7.51	0.18	0.020	9.06
vowel type: [ae]	0.24	0.16	1.48	0.24	0.16	1.48
vowel type: [eh]	-0.27	0.16	-1.65	-0.33	0.16	-2.000
vowel type: [ey]	-0.22	0.14	-1.57	-0.22	0.14	-1.56
vowel type: [ih]	-0.43	0.15	-2.86	-0.45	0.15	-2.98
vowel type: [iy]	0.036	0.15	0.25	0.065	0.15	0.44
vowel type: [ow]	-0.26	0.15	-1.76	-0.26	0.15	-1.77
vowel type: [uh]	0.23	0.22	1.05	0.21	0.22	0.96
vowel type: [uw]	-0.100	0.17	-0.600	-0.12	0.17	-0.75

Table 5.11: Summary of coefficients of the fixed-effects predictors for the vowel dispersion model before and after removing outliers. Before removing outliers, the model contains 9,656 tokens of 418 word types. After removing outliers, the model contains 9,477 tokens of 418 word types. "β" denotes the mean estimation of the coefficient; "S.E.' denotes the standard error in the estimation of the coefficient.



(a) Before removing outliers (b) After removing outliers

Figure 5.11: Q-Q plot of the residuals in the vowel dispersion model before and after removing 179 outliers.

5.3.2 Model summary

In this section, I will present a summary of the baseline model on K-dispersion[6], with an emphasis on neighborhood characteristics and phonotactic probability.

5.3.2.1 Model fit and random effects

The log likelihood of the baseline model is -10803. With all random and fixed effects, the proportion of variance explained by the model (R^2) is 0.433. But if we remove all the fixed-effects predictors and only retain the two random terms, the model can still achieve an R^2 of 0.426. In other words, including fixed-effects predictors does not contribute much to the percentage of explained variation.

However, the lack of change in R^2 does not mean that the fixed-effects terms are not effective in the model at all. As shown in Table 5.12, adding these terms reduces the standard deviation of by-word adjustment from 0.71 to 0.62 (i.e. 14% decrease) and the standard deviation of by-speaker adjustment from 0.29 to 0.27 (i.e. 6.9% decrease). The difference between word and speaker is expected, since several of the fixed effects are based on properties that vary by word. When these properties are modeled through fixed effects, random by-word adjustments do not need to model them any more.

Model	**Random effect**	**Variance**	**s.d.**
Baseline model	`word`	0.38	0.62
	`speaker`	0.073	0.27
	`residual`	0.52	0.72
No fixed-effects terms	`word`	0.50	0.71
	`speaker`	0.086	0.29
	`residual`	0.52	0.72

Table 5.12: Comparing random effects in the baseline vowel dispersion model (upper) and a model with random effects only (lower).

5.3.2.2 Fixed effects

A potential problem in a mixed-effect model is collinearity, which is mainly caused by high correlations among numerical predictors. As shown in Table 5.8 and 5.10, pair-wise correlations among the numerical predictor variables are in general quite low, except for the correlations among neighborhood density and the two phonotactic probabilities (r>0.5 in all three cases). Overall collinearity of the model is usually assessed by a parameter called condition number, which can be generated by the `collin.fnc()` function in the **languageR**

[6]It should be noted that the baseline model still contains non-significant predictors, e.g. phonotactic probability. As a result, the estimates also reflect the model's attempt to incorporate insignificant predictors.

package (Baayen 2009). The condition number in the baseline model is 2.99, which is small enough that no collinearity problem needs to be worried about.

Table 5.13 (repeated from Table 5.11) summarizes the fixed-effects terms in the baseline model. As introduced in Chapter 3, given the current data size, an absolute t value greater than 2.5 will roughly correspond to a p value smaller than 0.01. Thus I will use $|t|>2.5$ as a working criterion for statistical significance for now. More sophisticated tests of t values will be presented in §5.3.3.

As shown in Table 5.13, all predictors in the baseline model, except for phonotactic probability (phoneme: β=0.056, t=0.33; biphone: β=0.051, t=0.55), are associated with some absolute t values of at least 2.45. If the variable is categorical, at least one level has a t value greater than 2.5.

Predictor	$\boldsymbol{\beta}$	**S. E.**	t **value**
(intercept)	0.91	0.16	5.78
neighborhood density	-0.018	0.0072	-2.45
neighbor frequency	0.52	0.16	3.13
bigram probability (before)	-0.020	0.0050	-3.96
manner of consonant (after): NASAL	-0.35	0.13	-2.77
manner of consonant (after): OBS	-0.45	0.100	-4.43
phonotactic probability (phoneme)	0.0051	0.17	0.029
phonotactic probability (biphone)	0.020	0.097	0.21
place of consonant (before): CORONAL	-0.45	0.11	-4.18
place of consonant (before): LABIAL	-0.24	0.11	-2.25
speaker sex: M	-0.27	0.088	-3.07
speech rate (after)	-0.11	0.025	-4.25
vowel duration	0.18	0.020	9.06
vowel type: [ae]	0.24	0.16	1.48
vowel type: [eh]	-0.33	0.16	-2.000
vowel type: [ey]	-0.22	0.14	-1.56
vowel type: [ih]	-0.45	0.15	-2.98
vowel type: [iy]	0.065	0.15	0.44
vowel type: [ow]	-0.26	0.15	-1.77
vowel type: [uh]	0.21	0.22	0.96
vowel type: [uw]	-0.12	0.17	-0.75

Table 5.13: Summary of coefficients of the fixed-effects predictors in the baseline model. The model is based on 9,477 tokens of 418 word types. "β" denotes the mean estimation of the coefficient; "S. E." denotes the standard error in the estimation of the coefficient.

Among the significant predictors, neighborhood density is associated with a negative coefficient (-0.018), a relatively small standard error (0.0072) and a relatively large t value

(-2.45). Taken together, the statistics suggest that a tendency exists for words from dense neighborhoods to be realized with more *centralized* vowels, which is consistent with the shortening effect of high density on word duration, as found in the first corpus study. By contrast, neighbor frequency, which has no effect on word duration, is associated with a positive coefficient (0.52), a small standard error (0.16) and a large t value (3.13) in the vowel model, suggesting that words with high-frequency neighbors have more *dispersed* vowels than those with low-frequency neighbors. It should be noted that the opposite directions of the neighborhood density effect and the neighbor frequency effect on vowel dispersion are not predicted by either the speaker-oriented hypothesis or the listener-oriented hypothesis.

Table 5.13 also indicates a few other significant trends. Everything else being equal, a word token tends to have a more dispersed vowel if (a) the word is less predictable from the preceding word (`bigram probability (before)`); (b) the final consonant is an approximant as opposed to nasal or obstruent (`manner of consonant (after)`); (c) the initial consonant is back (mostly velar) as opposed to coronal or labial (`place of consonant (after)`, though the difference between BACK and LABIAL is marginal); (d) the word token is produced by a female speaker as opposed to a male speaker (`speaker sex`); (e) the word token is followed by slow speech (`speech rate (after)`); (f) the vowel token is produced with long duration (`vowel duration`); (g) the target vowel is [aa] as opposed to [ih] (`vowel type`). The trends regarding bigram probability, speaker sex, speech rate, vowel duration and consonant features are all expected, given previous findings in the literature.

5.3.2.3 Partial effects

A partial effect refers to the effect of a certain predictor when other variables in the model are controlled. Relatedly, one can also calculate the predicted range of a predictor, which is the the range of variation in the outcome variable by only varying the critical predictor and keeping other predictors at the mean levels. Examining partial effects and predicted ranges will give us some idea of the size of the impact of certain predictors, beyond other factors in the model.

Figure 5.12 plots all partial effects in the baseline model, produced by the `plotLMER.fnc` function in the **languageR** package (Baayen 2009). The solid lines show the predicted partial effects and the broken lines show the MCMC-based confidence intervals (see the following section on model evaluation).

As shown in the subplot for neighborhood density, there is a clear negative relationship between density and K-dispersion when other factors are controlled. The confidence intervals around the predicted line indicate that the relationship is quite linear throughout the range of the data. As is (vaguely) shown in the subplot, the predicted range of neighborhood density is [0.57, 1.23]. On the other hand, the subplot for neighbor frequency shows that everything else being equal, higher neighbor frequency is associated with greater degree of vowel dispersion, and the effect is quite linear throughout the range. The predicted range of neighbor frequency is [0.49, 1.35]. In other words, the partial effects of the two neighborhood variables are similar in magnitude but opposite in direction.

Recall that in the word duration model, compared with neighborhood density, both frequency and speech rate have larger effects and bigram probability has a comparable size of effect (see Figure 4.6). In the current model, both speech rate (predicted range: [0.69, 1.10]) and bigram probability (predicted range: [0.81, 0.99]) have smaller effects than neighborhood variables, and frequency does not even make to the final model, due to the lack of significance. The reduced size of effects of frequency, predictability and speech rate in the vowel dispersion model is probably due to the fact that the model also includes vowel duration as a fixed-effect predictor, which is highly significant (β=0.18, t=9.06) and has a predicted range of [0.67, 1.21]. Therefore, if frequency, predictability and speech rate are more directly related with durational variation and/or duration-dependent vowel dispersion, then the predicting power of these variable will be significantly reduced when vowel duration is already controlled for. In that sense, current results also suggest that the effects of neighborhood characteristics have greater influence (than other predictors) on vowel dispersion beyond the part that is conditioned by duration.

Not surprisingly, in Figure 5.12, the predicted lines for phonotactic variables are basically flat, suggesting no effect of phonotactic probability when other predictors are kept constant, which is consistent with their small t values in model summary.

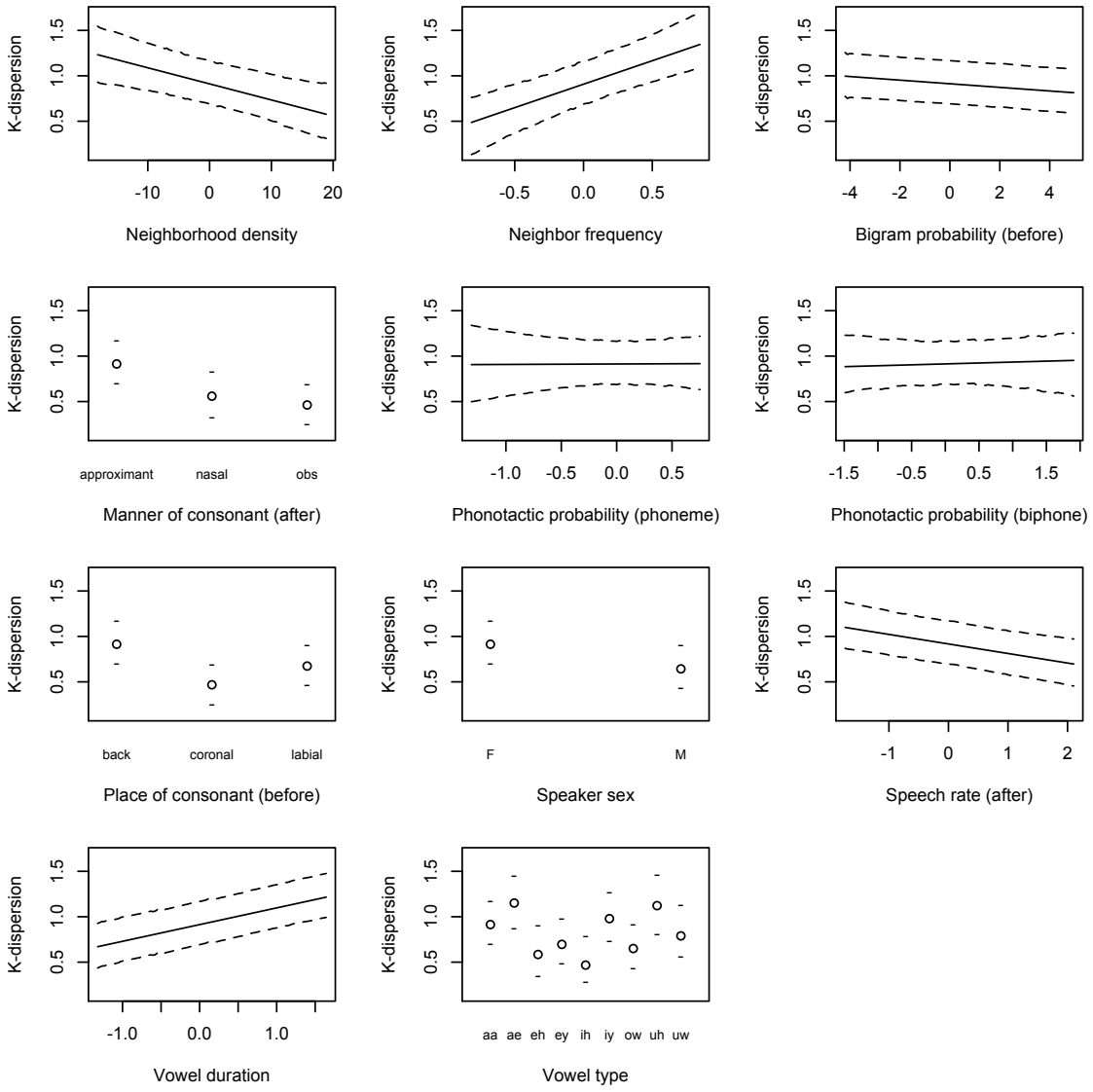


Figure 5.12: Partial effects in the baseline model. Y axis denotes log word duration in all subplots. X axis denotes individual predictor variable, after log transforming and centering, if any. The solid lines show the predicted lines for partial effects and the broken lines show the MCMC-base estimation of confidence intervals.

5.3.3 Model evaluation

Given the large size of the database and the subtlety of the effect under investigation, it is important to ensure that the observed effects are not spurious. Similar to the word duration model, a number of evaluation techniques (model comparison, MCMC-based testing and cross validation) were used to test the reliability and robustness of the model, especially regarding the effects of neighborhood characteristics and phonotactic probability.

5.3.3.1 Comparing models with and without neighborhood measures

Model comparison shows that when both neighborhood density and neighbor frequency are removed from the baseline model, model fit deteriorates significantly (Chisq=15.143, Chi Df=2, $p<0.001$). The same is true when only one neighborhood variable is removed (when removing density only, Chisq= 6.1936, Chi Df=1, $p=0.01$; when removing neighbor frequency only, Chisq=10.127, Chi Df = 1, $p=0.001$). These results suggest that both neighborhood predictors make important contribution to the fit of the model.

On the other hand, removing the two phonotactic measures does not cause a significant change in model fit (Chisq=0, Chi Df=2, $p=1$), which confirms their insignificance in the model.

5.3.3.2 Testing t values

The significance of the t values in the baseline model is further tested by a Markov chain Monte Carlo (MCMC) sampling technique. As introduced in Chapter 3, this technique generates a large number of samples of model parameters so that the significance of the coefficients can be evaluated based on their posterior distributions.

Table 5.14 summarizes the MCMC-based evaluation of model coefficients, generated by the `pvals.fnc()` function in the **lme4** package (Bates and Maechler 2010). Overall, model coefficients predicted by the MCMC method coincide with those reported in Table 5.13, and predictors with absolute t values around or above 2.5 in Table 5.13 are all associated with p values smaller than 0.05 in MCMC-based testing. In particular, the mean MCMC coefficient of neighborhood density is -0.018, which is exactly the same as the one given in Table 5.13, and the MCMC confidence interval of this parameter, [-0.0027, -0.0063], indicates that it is reliably smaller than zero ($p=0.0142$). Neighbor frequency has a mean MCMC coefficient of 0.52, which is also the same as the one on the summary table, and the confidence interval ([0.33, 0,81]) indicates that this parameter is reliably greater than zero ($p=0.0017$). The two phonotactic variables, on the other hand, both have MCMC confidence intervals that cross zero ([-0.24, 0.26] for phoneme probability; [-0.14, 0.14] for biphone probability) and p values that are almost 1.

Thus the significance of both neighborhood variables and the insignificance of phonotactic probability are both confirmed by the MCMC-based testing results.

Predictor	MCMC mean	HPD95 lower	HPD95 upper	Pr(>|t|)
(intercept)	0.91	0.68	1.16	0.0000
neighborhood density	-0.018	-0.027	-0.0063	0.0142
neighbor frequency	0.52	0.33	0.81	0.0017
bigram probability (before)	-0.020	-0.030	-0.011	0.0001
manner of consonant (after): NASAL	-0.35	-0.53	-0.16	0.0056
manner of consonant (after): OBS	-0.45	-0.62	-0.32	0.0000
phonotactic probability (phoneme)	0.0051	-0.24	0.26	0.9766
phonotactic probability (biphone)	0.020	-0.14	0.14	0.8358
place of consonant (before): CORONAL	-0.45	-0.61	-0.31	0.0000
place of consonant (before): LABIAL	-0.24	-0.41	-0.098	0.0247
speaker sex: M	-0.27	-0.43	-0.094	0.0021
speech rate (after)	-0.11	-0.16	-0.057	0.0000
vowel duration	0.18	0.15	0.23	0.0000
vowel type: [ae]	0.24	-0.0073	0.47	0.1392
vowel type: [eh]	-0.33	-0.55	-0.084	0.0458
vowel type: [ey]	-0.22	-0.39	0.00090	0.1190
vowel type: [ih]	-0.45	-0.62	-0.19	0.0029
vowel type: [iy]	0.065	-0.15	0.27	0.6575
vowel type: [ow]	-0.26	-0.47	-0.050	0.0761
vowel type: [uh]	0.21	-0.092	0.52	0.3365
vowel type: [uw]	-0.12	-0.34	0.15	0.4555

Table 5.14: Summary of coefficients estimated by MCMC sampling technique for the baseline model. "MCMC mean" denotes the mean estimation of coefficients, based on 10,000 Monte Carlo samples of the posterior distribution of model parameters; "HPD95 lower" and "HPD95 upper" denote the lower and upper bounds of Highest Posterior Density interval for 95% of the probability density; "Pr (>|t|)" denotes the probability based on the t-distribution with 9,477 degrees of freedom.

5.3.3.3 Testing model generalizability

As in the word duration study, two cross-validation tests, a type-based one and a token-based on, were used to evaluate model generalizability over the current set of words and tokens.

In the type-based test, in each iteration, I randomly extracted half of the words from the current word set (n=418) and fitted a model with the same structure as the baseline model using all tokens of the extracted words. The test consists of 100 iterations, and a summary of model coefficients over all iterations is given in Table 5.15. As shown in the table, neighborhood density has a mean coefficient of -0.017, which is close to the estimation of -0.018 in the baseline model. The 95% confidence interval of this parameter is from -0.031 to -0.0018, indicating it is reliably smaller than zero. Neighbor frequency has a mean coefficient of 0.500, which is also similar to the one in the baseline model (0.52), and a confidence interval of [0.14, 0.83] that is always greater than zero. Both phonotactic variables have confidence intervals across zero ([-0.36, 0.34] for phoneme probability; [-0.15, 0.30] for biphone probability), which, again, suggests that the phontactic effects are not reliable.

In the token-based cross-validation test, a similar procedure was applied. In each run, half of the tokens were randomly drawn from the complete dataset (n=9,656) and a model was fitted based on the extracted tokens. A summary of model coefficients over 100 runs is given in Table 5.16. As shown in the table, the coefficients of both neighborhood density (-0.018) and neighbor frequency (0.58) are similar to those in the baseline model, and the confidence intervals of both parameters ([-0.024, -0.0092] for density; [0.400, 0.71] for neighbor frequency) are consistently smaller or greater than zero, indicating statistical significance. Like in the type-based test, both phonotactic variables are associated with confidence intervals that cross zero ([-0.12, 0.19] for phoneme probability; [-0.090, 0.082] for biphone probability), which confirms the insignificance.

Taken together, results from the two cross-validation tests indicate that the findings from the baseline model, as shown in Table 5.13, are not likely to be an accident of the current dataset. Since the same trends can be reliably observed in randomly generated subsets of the current data, it gives us confidence to say that the current results may be generalized to unseen words and tokens of similar nature.

5.3.3.4 Summary of model evaluation results

To sum up, the baseline model reveals a negative effect of neighborhood density and a postive effect of neighbor frequency on vowel dispersion. No reliable effects of phonotactic probability have been found. The reliability and robustness of these findings is confirmed by all evaluation tests that have been applied. Everything else being equal, a high-density word tends to have a more *centralized* vowel than a low density word, but a word with high-frequency neighbors tends to have a more *dispersed* vowel than a word with low-frequency neighbors. The sizes of the neighborhood effects are greater than that of some other numerical predictors in the model, including bigram probability and speech rate.

Predictor	2.5%	50%	97.5%
(intercept)	0.64	0.92	1.27
neighborhood density	-0.031	-0.017	-0.0018
neighbor frequency	0.14	0.500	0.83
bigram probability (before)	-0.033	-0.020	-0.0049
manner of consonant (after): NASAL	-0.64	-0.33	-0.036
manner of consonant (after): OBS	-0.74	-0.43	-0.18
phonotactic probability (phoneme)	-0.36	0.025	0.34
phonotactic probability (biphone)	-0.15	0.0082	0.300
place of consonant (before): CORONAL	-0.76	-0.47	-0.22
place of consonant (before): LABIAL	-0.57	-0.26	-0.042
speaker sex: M	-0.34	-0.28	-0.200
speech rate (after)	-0.16	-0.11	-0.046
vowel duration	0.077	0.18	0.33
vowel type: [ae]	-0.099	0.24	0.500
vowel type: [eh]	-0.61	-0.33	-0.0089
vowel type: [ey]	-0.44	-0.23	0.040
vowel type: [ih]	-0.79	-0.46	-0.100
vowel type: [iy]	-0.22	0.059	0.33
vowel type: [ow]	-0.54	-0.28	-0.0012
vowel type: [uh]	-0.39	0.200	0.78
vowel type: [uw]	-0.47	-0.12	0.18

Table 5.15: Summary of type-based cross validation results for the baseline model. Type-based cross validation is based on 100 random subsets of half of the word types. In the table, "50%" denotes the mean value of the coefficient over 100 test models; "2.5%" and "97.5%" denote the lower and upper bounds of the 95% density area of the coefficient.

Predictor	2.5%	50%	97.5%
(intercept)	0.83	0.93	1.04
neighborhood density	-0.024	-0.018	-0.0092
neighbor frequency	0.400	0.58	0.71
bigram probability (before)	-0.028	-0.019	-0.010
manner of consonant (after): NASAL	-0.49	-0.36	-0.22
manner of consonant (after): OBS	-0.57	-0.48	-0.39
phonotactic probability (phoneme)	-0.12	0.0012	0.19
phonotactic probability (biphone)	-0.090	0.0078	0.082
place of consonant (before): CORONAL	-0.54	-0.45	-0.36
place of consonant (before): LABIAL	-0.34	-0.24	-0.14
speaker sex: M	-0.31	-0.27	-0.24
speech rate (after)	-0.15	-0.11	-0.060
vowel duration	0.14	0.18	0.21
vowel type: [ae]	0.076	0.22	0.38
vowel type: [eh]	-0.43	-0.300	-0.19
vowel type: [ey]	-0.33	-0.19	-0.076
vowel type: [ih]	-0.53	-0.41	-0.27
vowel type: [iy]	-0.069	0.064	0.16
vowel type: [ow]	-0.38	-0.26	-0.14
vowel type: [uh]	0.061	0.21	0.38
vowel type: [uw]	-0.19	-0.076	0.095

Table 5.16: Summary of token-based cross validation results for the baseline model. Token-based cross validation is based on 100 random subsets of half of the tokens. In the, "50%" denotes the mean value of the coefficient over 100 test models; "2.5%" and "97.5%" denote the lower and upper bounds of the 95% density area of the coefficient.

5.4 Alternative analyses

The preceding section presented several model evaluation tests, which all confirmed the reliability of the trends shown in the baseline model. However, the baseline model only represents one of the many possible ways for estimating neighborhood metrics and vowel dispersion. To find out whether the observed effects are a result of how these variables are estimated, I conducted two alternative analyses, one with frequency-weighted (instead of raw) neighborhood density and the other with CELEX-based (instead of HML-based) neighborhood measures. The results from these alternative analyses are presented in the following. To preview the results, frequency-weighted neighborhood density has no effect on vowel dispersion. In the CELEX model, neighborhood density has a significant centralizing effect, which is consistent with the baseline model, but the effect of neighbor frequency does not reach signficance. .

5.4.1 Using frequency-weighted neighborhood density

5.4.1.1 Calculating frequency-weighted neighborhood density

As discussed in the background chapter, the literature on phonological neighborhoods has exploited both raw neighborhood density and frequency-weighted density measures. An underlying assumption in doing so is that high neighborhood density and high neighbor frequency will have the same effects on word perception and production. However, in the baseline model, we have seen a different pattern. While high density is shown to have a centralizing effect on vowels, high neighbor frequency seems to be associated with greater vowel dispersion. Thus, the effects of neighborhood density and neighbor frequency on vowel dispersion are clearly *not* in the same direction, but testing the model with a frequency-weighted density measure can still be rewarding, because it will help reveal the relative strengths of the two neighborhood effects.

Frequency-weighted neighborhood density (FWND) can be computed as the sum of the frequencies of all neighbors (i.e. $\sum f_i \cdot 1$), which equals the multiplicational product of raw density and average neighbor frequency (i.e. $\bar{f} \cdot ND$). In the complete vowel database[7] (n=9,656), FWND ranges from 4.43 to 91.99 (in log frequency), with a median value of 45.91 and a mean value of 46.86 (s.d. = 17.24). Figure 5.13 plots the distribution of FWND over word types in the dataset. As reported in the word duration study, FWND is highly correlated with raw density (r=0.95) and also relatively correlated with neighbor frequency (r=0.42).

[7]This refers to the vowel dataset before removing the 179 outliers

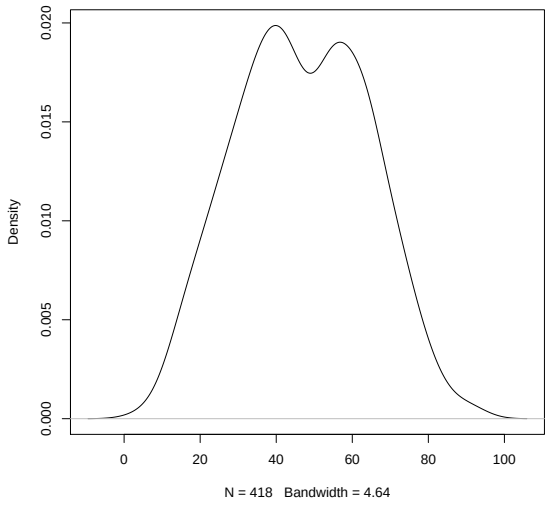


Figure 5.13: Distribution of FWND in the word set (n=418).

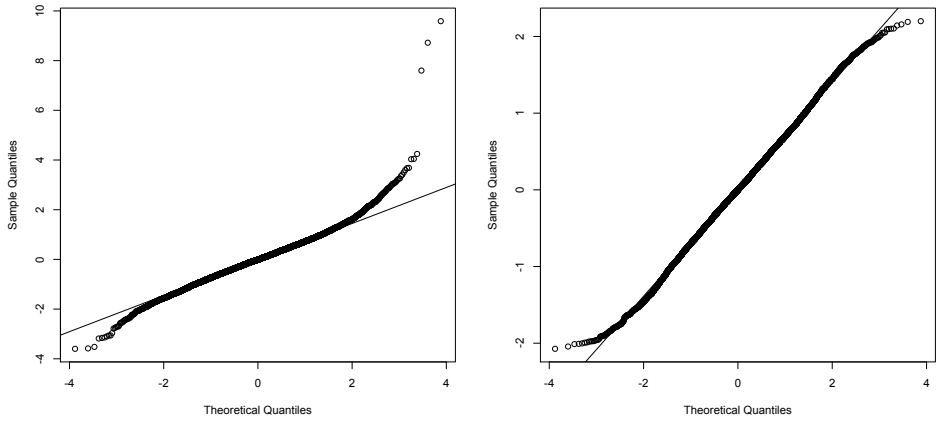
5.4.1.2 Modeling with frequency-weighted neighborhood density

The model with frequency-weighted neighborhood density (i.e. the "FWND model"[8]) was built with the same procedure as the baseline model. An initial model was first built, using the complete vowel database. All predictor variables were included in the initial model, except for neighborhood density and neighbor frequency, which were replaced by FWND. In the trimming process, 14 predictors were removed (1^{st} round: bigram probability (after), disfluency (before), familiarity, frequency, previous mention, speech rate (before), voicing of consonant (after); 2^{nd} round: disfluency (after), manner of consonant (before), orthographic length, part of speech, place of consonant (after), speaker age, voicing of consonant (before)), as well as 179 outlier data points. The removal of outliers significantly improves model fit (Figure 5.14), without changing model coefficients (Table 5.17).

Overall, the model suggests that FWND is *not* a significant predictor for K-dispersion (β = -0.0025, t = -0.78). Neither are the two phonotactic variables (phoneme: β = -0.14, t = -0.83; biphone: β = 0.011, t = 0.11). But the significant effects in the FWND model, concerning bigram probability, manner of following consonant, place of preceding consonant, speaker sex, following speech rate, vowel duration and vowel type, are consistent with the baseline model.

The insignificance of FWND and phonotactic probability is confirmed by model comparison (p>0.4 in all cases), MCMC-based t tests and cross validation tests. Results from the

[8]In this chapter, without further specification, the term "FWND model" refers to the FWND model for **vowel dispersion**.



(a) Before removing outliers (b) After removing outliers

Figure 5.14: Q-Q plot of the residuals in the FWND Model, before and after removing 179 outliers.

latter two tests are reported in Table 5.18. Overall, the results suggest that when neighborhood density and neighbor frequency are combined, the resulting variable has no significant effect on vowel dispersion.

Predictor	Before removing outliers			After removing outliers		
	β	S. E.	t value	β	S. E.	t value
(intercept)	0.900	0.16	5.64	0.93	0.16	5.81
FWND	-0.0026	0.0032	-0.82	-0.0025	0.0032	-0.78
bigram probability (before)	-0.019	0.0056	-3.35	-0.020	0.0050	-3.96
manner of consonant (after): NASAL	-0.31	0.13	-2.39	-0.33	0.13	-2.53
manner of consonant (after): OBS	-0.47	0.100	-4.53	-0.49	0.100	-4.77
phonotactic probability (phoneme)	-0.12	0.17	-0.68	-0.14	0.17	-0.83
phonotactic probability (biphone)	0.0093	0.099	0.094	0.011	0.099	0.11
place of consonant (before): CORONAL	-0.49	0.11	-4.50	-0.500	0.11	-4.65
place of consonant (before): LABIAL	-0.23	0.11	-2.13	-0.23	0.11	-2.10
speaker sex: M	-0.25	0.089	-2.85	-0.27	0.088	-3.07
speech rate (after)	-0.12	0.028	-4.26	-0.11	0.025	-4.25
vowel duration	0.17	0.023	7.54	0.18	0.020	9.09
vowel type: [ae]	0.21	0.17	1.24	0.200	0.16	1.24
vowel type: [eh]	-0.11	0.16	-0.68	-0.17	0.16	-1.02
vowel type: [ey]	-0.25	0.14	-1.79	-0.25	0.14	-1.78
vowel type: [ih]	-0.41	0.15	-2.67	-0.42	0.15	-2.76
vowel type: [iy]	0.035	0.15	0.23	0.064	0.15	0.43
vowel type: [ow]	-0.33	0.15	-2.18	-0.33	0.15	-2.17
vowel type: [uh]	0.34	0.22	1.57	0.33	0.22	1.49
vowel type: [uw]	-0.11	0.17	-0.62	-0.13	0.17	-0.76

Table 5.17: Summary of coefficients of the fixed-effects predictors for the FWND model, before and after removing outliers. Before removing outliers, the model contains 9,656 tokens of 418 word types. After removing outliers, the model contains 9,477 tokens of 418 word types. "β" denotes the mean estimation of the coefficient; "S.E.' denotes the standard error in the estimation of the coefficient.

Predictor	MCMC mean	HPD95 lower	HPD95 upper	Pr(>|t|)
FWND	0	-0.0025	-0.0019	0.0028
phonotactic probability (phoneme)	-0.14	-0.400	0.100	0.4039
phonotactic probability (biphone)	0.011	-0.14	0.14	0.9131

(a) MCMC-based evaluation

Predictor	2.5%	50%	97.5%
FWND	-0.010	-0.0031	0.0036
phonotactic probability (phoneme)	-0.53	-0.11	0.31
phonotactic probability (biphone)	-0.21	0.011	0.200

(b) Type-based cross validation

Predictor	2.5%	50%	97.5%
FWND	-0.0044	-0.0016	0.0012
phonotactic probability (phoneme)	-0.300	-0.13	-0.019
phonotactic probability (biphone)	-0.091	-0.0039	0.077

(c) Token-based cross validation

Table 5.18: Summary of model evaluation results for the FWND model. MCMC-based evaluation was based on 10,000 Monte Carlo samples of posterior distribution of the model coefficients. Type-based cross validation was conducted with 100 random subsets of the data, each containing half of the word types. Reported model coefficients are summarized over all test samples. Token-based cross validation was based on 100 random subsets of the data, each containing half of the word tokens. Reported model coefficients are summarized over all test samples. Only results regarding FWND and phonotactic probability are shown in the table.

Variable	Min	Max	Median	Mean	s.d.
CELEX density	6	43	26	25.42	8.26
CELEX neighbor frequency	3.54	7.65	5.34	5.43	0.52
CELEX FWND	29.41	266.60	141.10	137.20	43.78

Table 5.19: Summary statistics for CELEX density, CELEX neighbor frequency and CELEX FWND in the vowel database, based on 9,656 word tokens.

5.4.2 Using CELEX-based neighborhood measures

5.4.2.1 Calculating CELEX-based neighborhood measures

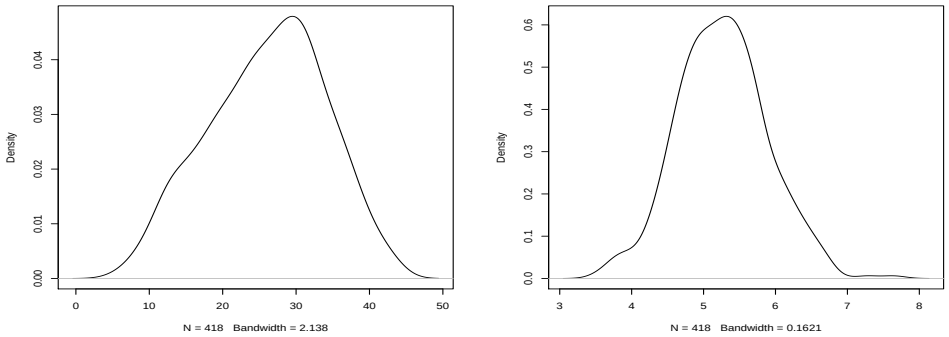
The second alternative analysis uses neighborhood measures computed based on citation forms from the CMU pronunciation dictionary and word form frequency from the CELEX lexical database (Baayen et al. 1993). The CELEX frequency is in turn derived from the 17.9-million-word COBUILD corpus, which is both larger in size and more recent in time than the Brown corpus (∼1 million word), which is the source of the frequency measures in HML.

Furthermore, when calculating the CELEX-based neighborhood measures, I adopted a frequency threshold for neighborhood membership. Only words with a log CELEX frequency of 4 or higher (i.e. 54 or more occurrences in the COBUILD corpus) can be considered as a potential neighbor. This frequency threshold is chosen to roughly represent a high familiarity rating (e.g. 6 or higher on a 7-point scale), based on the correlations between familiarity ratings from HML and word frequency in CELEX. (For more detail, please refer to §4.4.2.)

Altogether three CELEX-based neighborhood measures were computed: raw density ("CELEX density"), mean log neighbor frequency ("CELEX neighbor frequency"), and log frequency-weighted density ("CELEX FWND"). Table 5.19 shows the summary statistics of these variables in the complete vowel database (n=9,656). Figure 5.15 plots the distribution of the variables over the word set (n=418). Both CELEX density and CELEX neighbor frequency are correlated with the corresponding HML metric over the word set (r=0.76 for density; r=0.74 for neighbor frequency). Similar to the HML metrics, CELEX FWND is extremely highly correlated with CELEX density (r=0.94), and sizably correlated with CELEX neighbor frequency (r=0.37).

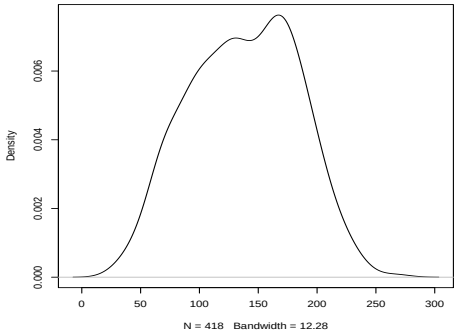
5.4.2.2 Modeling with CELEX-based neighborhood measures

Model construction was carried out with the same procedure as described before. Thirteen variables were eliminated during the trimming process (1_{st} round: bigram probability (after), disfluency (before), familiarity, frequency, manner of consonant (before), previous mention, speech rate (before), voicing of consonant (after); 2_{nd} round: disfluency (after),



(a) CELEX density

(b) CELEX neighbor frequency



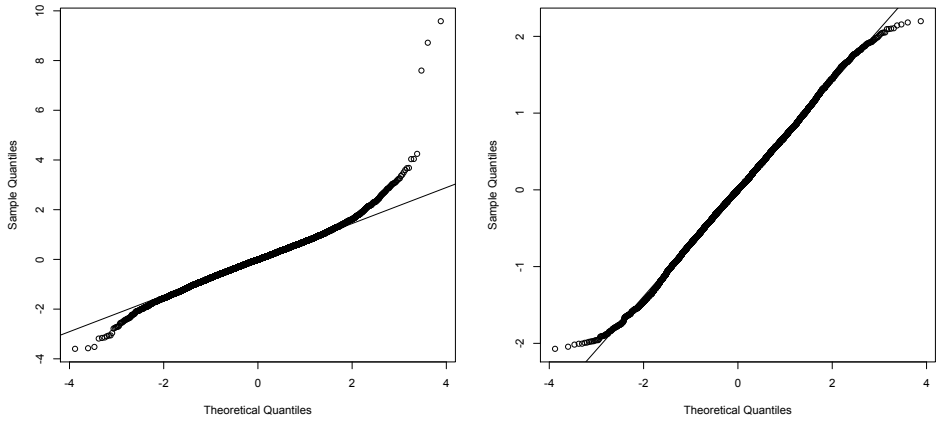
(c) CELEX FWND

Figure 5.15: Distribution of CELEX-based neighborhood measures in the vowel database. For all variables, probability density functions over the word set (n=418) are shown.

orthographic length, part of speech, place of consonant (after), speaker age). In addition, 179 outliers were removed from the data, which improves model fit (Figure 5.16) but does not change model coefficients (Table 5.20).

The final model (i.e. the "CELEX model"[9]) contained 14 fixed-effects predictors, based on 9,477 tokens of 418 words. As shown in Table 5.20, CELEX density has a negative effect on vowel dispersion (β=-0.017, t= -3.04), which is consistent with the baseline model, but CELEX neighbor frequency has no effect (β=0.013, t= 0.2). Phonotactic probability is not significant ($|t|$<1 for both predictors), either. Model comparison confirms the significance of CELEX density (Chisq=9.46 ; df=1; p=0.002) and the insignificance of CELEX neighbor frequency as well as phonotactic probability (p >0.3 in all cases). These trends are also at-

[9]In this chapter, without further specification, the term "CELEX model" refers to the CELEX model for **vowel dispersion**.



(a) Before removing outliers (b) After removing outliers

Figure 5.16: Q-Q plot of the residuals in the CELEX model before and after removing 179 outliers.

tested by MCMC-based evaluation and cross-validation tests (see Table 5.21). The predicted range of CELEX density is [0.50, 1.11], which is slightly smaller than that in the baseline model ([0.57, 1.23]).

Furthermore, if we model with CELEX-based frequency-weighted density (CELEX FWND), the combined density measure generates an overall centralizing effect (β=-0.0034, t=-3.20) in the resulting model, which is expected given the centralizing effect of density and null effect of neighbor frequency in the CELEX model.

To sum up, the CELEX model finds an overall centralizing effect of neighborhood density, but no reliable effects of neighbor frequency or phonotactic probability.

Predictor	Before removing outliers			After removing outliers		
	β	S. E.	t value	β	S. E.	t value
(intercept)	0.76	0.17	4.49	0.79	0.17	4.66
CELEX density	-0.016	0.0055	-2.92	-0.017	0.0054	-3.04
CELEX neighbor frequency	0.0022	0.068	0.033	0.013	0.067	0.200
bigram probability (before)	-0.018	0.0056	-3.30	-0.020	0.0050	-3.92
manner of consonant (after): NASAL	-0.29	0.13	-2.28	-0.31	0.13	-2.41
manner of consonant (after): OBS	-0.44	0.100	-4.28	-0.46	0.100	-4.50
phonotactic probability (phoneme)	-0.14	0.17	-0.79	-0.15	0.17	-0.88
phonotactic probability (biphone)	0.063	0.096	0.66	0.066	0.096	0.69
place of consonant (before): CORONAL	-0.43	0.11	-3.89	-0.44	0.11	-4.01
place of consonant (before): LABIAL	-0.18	0.11	-1.60	-0.17	0.11	-1.56
speaker sex: M	-0.25	0.089	-2.85	-0.27	0.088	-3.07
speech rate (after)	-0.12	0.028	-4.26	-0.11	0.025	-4.25
voicing of consonant (before): VOICELESS	0.16	0.080	2.06	0.17	0.079	2.10
vowel duration	0.17	0.023	7.57	0.18	0.020	9.12
vowel type: [ae]	0.200	0.16	1.21	0.19	0.16	1.19
vowel type: [eh]	-0.16	0.17	-0.96	-0.22	0.17	-1.35
vowel type: [ey]	-0.22	0.14	-1.64	-0.22	0.14	-1.63
vowel type: [ih]	-0.400	0.15	-2.66	-0.42	0.15	-2.78
vowel type: [iy]	0.038	0.15	0.26	0.066	0.15	0.45
vowel type: [ow]	-0.37	0.15	-2.47	-0.37	0.15	-2.48
vowel type: [uh]	0.26	0.22	1.22	0.24	0.22	1.11
vowel type: [uw]	-0.080	0.17	-0.47	-0.100	0.17	-0.600

Table 5.20: Summary of coefficients of the fixed-effects predictors for the CELEX model before and after removing outliers. Before removing outliers, the model contains 9,656 tokens of 418 word types. After removing outliers, the model contains 9,477 tokens of 418 word types. "β" denotes the mean estimation of the coefficient; "S.E.' denotes the standard error in the estimation of the coefficient.

Predictor	MCMC mean	HPD95 lower	HPD95 upper	Pr(>|t|)
CELEX density	-0.016	-0.023	-0.0073	0.0024
CELEX neighbor frequency	0.013	-0.080	0.12	0.8427
phonotactic probability (phoneme)	-0.15	-0.400	0.096	0.3798
phonotactic probability (biphone)	0.066	-0.086	0.19	0.4916

(a) MCMC-based evaluation

Predictor	2.5%	50%	97.5%
CELEX density	-0.029	-0.017	-0.0089
CELEX neighbor frequency	-0.15	0.010	0.13
phonotactic probability (phoneme)	-0.46	-0.10	0.19
phonotactic probability (biphone)	-0.11	0.060	0.28

(b) Type-based cross validation

Predictor	2.5%	50%	97.5%
CELEX density	-0.018	-0.014	-0.0092
CELEX neighbor frequency	-0.041	0.026	0.083
phonotactic probability (phoneme)	-0.27	-0.16	0.0053
phonotactic probability (biphone)	-0.022	0.049	0.12

(c) Token-based cross validation

Table 5.21: Summary of model evaluation results for the CELEX model. MCMC-based evaluation was based on 10,000 Monte Carlo samples of posterior distribution of the model coefficients. Type-based cross validation was conducted with 100 random subsets of the data, each containing half of the word types. Reported model coefficients are summarized over all test samples. Token-based cross validation was based on 100 random subsets of the data, each containing half of the word tokens. Reported model coefficients are summarized over all test samples. Only results regarding CELEX density, CELEX neighbor frequency and phonotactic probability are shown in the table.

5.5 Chapter discussion

5.5.1 Potential confounding factors

Generally speaking, there are two potential confounding factors for neighborhood effects on vowel dispersion. The first factor is vowel duration. As discussed before, there is a default correlation between vowel duration and vowel dispersion, due to duration-dependent vowel undershoot/overshoot (Moon and Lindblom 1994). In the current study, vowel duration is statistically controlled for by being included in the model as a fixed-effects predictor. In that sense, any observed neighborhood effect in the current study is *beyond* the part of variation in vowel dispersion that can be attributed to durational variation. Thus, the current findings of neighborhood effects are not likely to be confounded by vowel duration.

Another potential confounding factor is phonotactic probability. Phonotactic probability, measured by either phoneme probability or biphone probability, has a sizable correlation (r>0.5 in both cases) with neighborhood density in the current vowel database. However, in the baseline model, as well as the FWND model and the CELEX model, no significant effect of phonotactic probability on vowel dispersion has been observed. Neighborhood density (in raw value) has a significant centralizing effect in both the baseline model and the CELEX model. This suggests that neighborhood density is at least a better predictor for degree of vowel dispersion than phonotactic probability is. In order to test whether the effect of neighborhood density is dependent on the presence of phonotactic probability, I remove the two phonotactic measures from the baseline model. In the resulting model, neighborhood density still has a significant centralizing effect on vowel production, with largely unchanged parameters (β = -0.017; t = -3.05), which suggests that the effect of neighborhood density is *not* affected by the removal of phonotactic probability. On the other hand, if neighborhood density is removed from the baseline model, the two phonotactic measures still fail to reach significance ($|t|$<1.2 in both cases).

Taken together, results of model testing show that the centralizing effect of neighborhood density exists in the model, regardless of the presence or absence of phonotactic probability, but phonotactic probability is *not* a significant predictor for vowel dispersion, whether or not neighborhood density is included in the model. That being said, the observed effect of neighborhood density is not likely to be confounded by phonotactic probability.

To sum up, the significant effects of neighborhood characteristics in both the baseline model and the model with front vowels are not likely to be an artifact of the presence or absence of phonotactic probability in the model.

5.5.2 Neighborhood density v.s. neighbor frequency

In the current study, we have seen that neighborhood density has an overall centralizing effect on vowel production, while neighbor frequency seems to have a dispersing effect. Not surprisingly, when the two variables are combined in a frequency-weighted density measure (in the FWND model), no effect is observed. However, it is worth noting that this null effect

should not be considered as suggesting that the neighborhood effects in the baseline model are spurious. From a modeling perspective, neighborhood density and neighbor frequency are largely independent (r=0.13), therefore it is not likely for these variables to affect each other in a linear model. In fact, when either neighborhood density or neighbor frequency is removed from the baseline model, the remaining neighborhood variable is still significant, with largely unchanged coefficients. Therefore, a more probable explanation for the null effect of FWND is that words with high FWND (high density, high neighbor frequency) and those with low FWND (low density, low neighbor frequency) have similarly dispersed vowels because the effects of density and neighbor frequency cancel each other out. In other words, the null effect of FWND simply indicates that the centralizing effect of density and the dispersing effect of neighbor frequency are of similar size, a point that is corroborated by the comparable predicted ranges of density ([0.57, 1.23]) and neighbor frequency ([0.49, 1.35]) in the baseline model.

Although the effects of neighborhood density and neighbor frequency seem to be similar in size in the baseline model, the two effects are not equally stable across dictionaries. When CELEX-based neighborhood measures are used, neighborhood density still exhibits a significant centralizing effect, but the effect of neighbor frequency fails to reach significance. Thus, the reliability of the dispersing effect of neighbor frequency remains to be tested in future research.

5.5.3 Individual differences

Last but not least, we would also like to gauge the amount of individual speaker differences regarding the neighborhood effects in the vowel dispersion model. To do this, I fitted two test models, both on the basis of the baseline model. The first test model (i.e. the "speaker crossed with density" model) contains an additional random effect of speaker regarding the slope of the effect of neighborhood density, whereas the second model (i.e. the "speaker crossed with neighbor frequency" model) contains an additional term of speaker regarding the slope of the effect of neighbor frequency. Thus, in addition to by-speaker adjustments to the intercept, these models also assign by-speaker adjustments to the coefficients of neighborhood effects.

In the "speaker crossed with density" model, neighborhood density has a general coefficient of -0.017, highly similar to the one in the baseline model (-0.018), and a marginal t value of -2.36. The standard deviation of by-speaker adjustment for density coefficient is relatively small (0.008), indicating that the individualized density coefficients, which are the sum of the overall coefficient and individual adjustment, are mostly smaller than zero[10]. Model comparison with the baseline model indicates that the addition of the random term is

[10]In the "speaker crossed with density" model, the slope of the density effect for speaker S_i can be expressed as $\beta_{ND} + \gamma_{ND,i}$, where β_{ND} is the model coefficient for neighborhood density (i.e. -0.017), and $\gamma_{ND,i}$ is the adjustment in the slope of neighborhood density for speaker S_i. Since γ_{ND} is normally distributed around zero, with a standard deviation of about 0.008, it is safe to say that the sum of β_{ND} and $\gamma_{ND,i}$ is probably smaller than zero.

significant (Chisq=22.72, Chi Df=2, $p<0.001$). Figure 5.17 plots the individualized partial effects of neighborhood density in the model against scatterplots of K-dispersion and density, generated by a customized panel function (Baayen 2008) for `xyplot()` in the **lattice** package (Sarkar 2008). As shown in the plot, most speakers have a decreasing K-dispersion when neighborhood density goes up, suggesting a centralizing effect of density. However, the tendency is very weak in some speakers (e.g. s01 and s12), and in one speaker, s03, the effect is even in the opposite direction.

In the "speaker crossed with neighbor frequency" model, neighbor frequency has an overall coefficient of 0.53, highly similar to the one in the baseline model (0.52), and a marginal t value of 3.01. The standard deviation of by-speaker adjustment for density coefficient is 0.26, which indicates that individualized neighbor frequency coefficients should mostly be greater than zero. Model comparison also suggests that the addition of the random term regarding neighbor frequency is significant (Chisq=27.90, Chi Df=2, $p<0.001$). Figure 5.18 plots the individualized partial effects of neighbor frequency in the model, against scatterplots of K-dispersion and neighbor frequency. As shown in the plot, most speakers exhibit a clear dispersing effect of neighbor frequency on K-dispersion. However, the trend is definitely stronger in some speakers (e.g. s25 and s40) than in others (e.g. s02 and s08), which explains the gain in model fit when neighbor frequency effects are individually fitted.

To sum up, the centralizing effect of neighborhood density and the dispersing effect of neighbor frequency are present in almost all of the speakers. Adding individualized adjustment for neighborhood effects improves model fit, but do not change the direction of the neighborhood effects. Thus, it is safe to conclude that the observed neighborhood effects in the baseline model hold across speakers in the Buckeye corpus and can probably be generalized to other speakers.

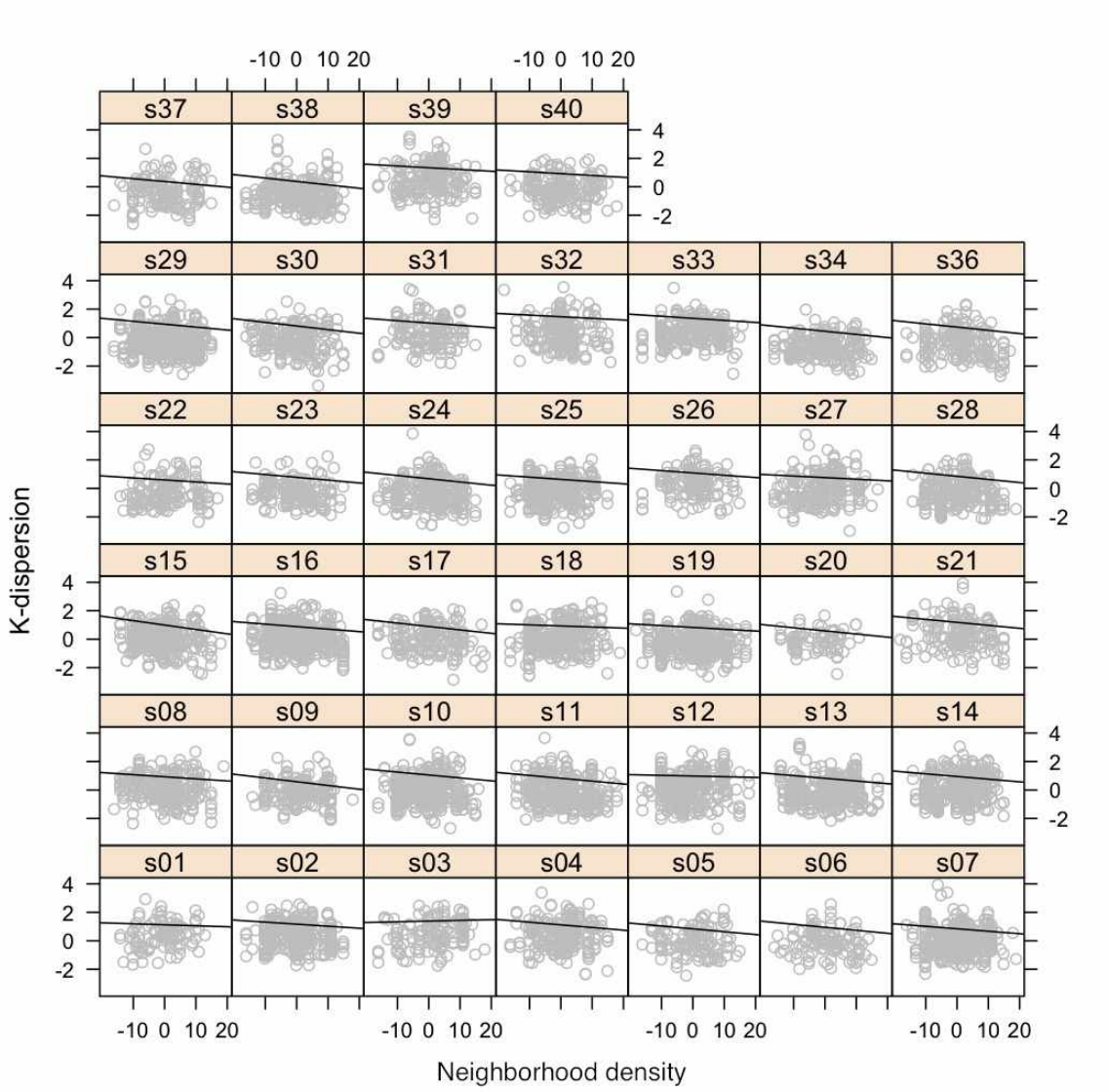


Figure 5.17: Individualized partial effects of neighborhood density on vowel dispersion. The grey circles in the background are scatterplots of K-dispersion and neighborhood density for each speaker (based on the 9,477-token dataset for the baseline model). Y axis denotes K-dispersion; X axis denotes centered neighborhood density from HML. The solid lines represent the individualized partial effects of neighborhood density, based on a model with a random speaker effect regarding the coefficient of neighborhood density. Other aspects of the model are the same as the baseline model.

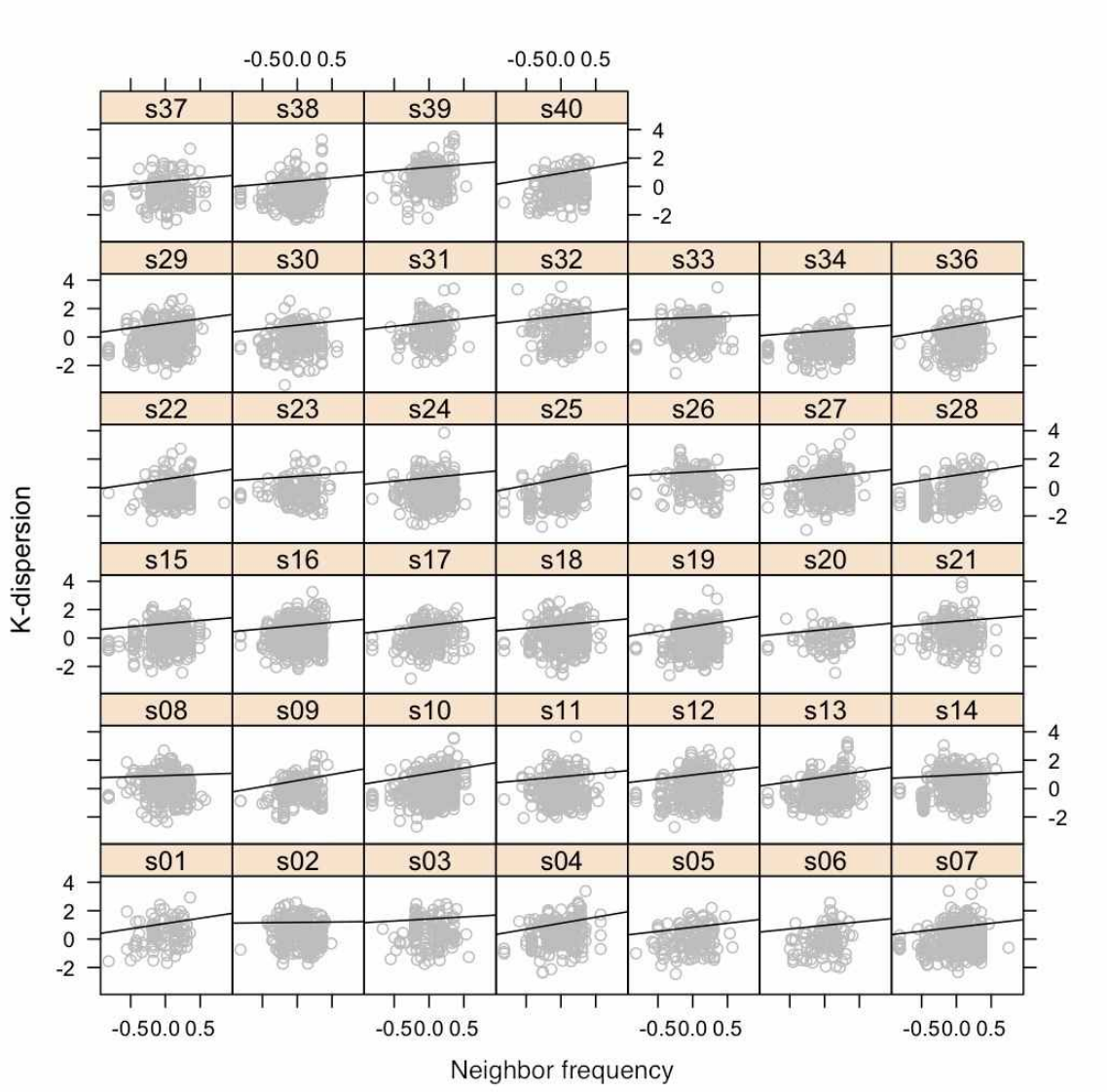


Figure 5.18: Individualized partial effects of neighbor frequency on vowel dispersion. The grey circles in the background are scatterplots of K-dispersion and neighbor frequency for each speaker (based on the 9,477-token dataset for the baseline model). Y axis denotes K-dispersion; X axis denotes centered neighbor frequency from HML. The solid lines represent the individualized partial effects of neighbor frequency, based on a model with a random speaker effect regarding the coefficient of neighbor frequency. Other aspects of the model are the same as the baseline model.

5.5.4 Summary

In this study, I started with two sets of hypotheses for neighborhood effects on vowel production. The speaker-oriented hypothesis predicts vowel *centralization* in words from high-density neighborhoods and words with high-frequency neighbors. Contrarily, the listener-oriented hypothesis predicts vowel *dispersion* in words from high-density neighborhoods and words with high-frequency neighbors.

Results from the corpus study reveal a centralizing effect of neighborhood density, as the vowels in high-density words are closer to the center of vowel space than the vowels in low-density words. The effect remains significant when neighborhood density is estimated from an alternative (i.e. CELEX) dictionary. Neighbor frequency, on the other hand, seems to have a dispersing effect on vowel production, suggesting that the vowels in words with high-frequency neighbors tend to be farther away from the center of vowel space than the vowels in words with low-frequency neighbors. However, the reliability of the neighbor frequency effect is not fully attested, as the significance of this effect seems to go away when CELEX-based neighbor frequency is used.

Thus, it is safe to say that the most robust finding of the vowel study is that high neighborhood density has a centralizing effect on vowel production, which provides support for the speaker-oriented hypothesis. One may argue that reduced articulatory effort is more directly related with vowel assimilation, which is not necessarily the same as vowel centralization. Albeit a valid point, it should not undermine the ground for the current study, because vowel centralization is a well-established feature of phonetic reduction, and the aggregate result of vowel-to-vowel coarticulation may be similar to vowel centralization.

In the next chapter, I will discuss the general theoretical implications of findings from both corpus studies.

Chapter 6

General Discussion

6.1 Summary of findings from the corpus studies

The major goal of this research is to investigate the effects of neighborhood structure on pronunciation variation in conversational speech. For this purpose, two corpus studies were carried out on the pronunciation of CVC monomorphemic content words in the Buckeye corpus. The first study investigates neighborhood effects on word duration and the second study investigates neighborhood effects on vowel dispersion. In both studies, a series of mixed-effect linear regression models were built, with speaker and word as two random terms and a wide range of factors including neighborhood density and neighbor frequency as fixed-effects predictors. Both neighborhood variables in the main models come from the Hoosier mental lexicon (HML; Nusbaum et al. 1984).

As shown in Chapter 4, neighborhood density has a negative effect on word duration. Everything else being equal, words from dense neighborhoods are realized with *shorter* durations than words from sparse neighborhoods. Importantly, this effect is *not* confounded by phonotactic probability, i.e. the frequency of phonological patterns in the word, which has a default correlation with neighborhood density. Results from model evaluation confirm that the shortening effect of density reliably exists in random subsets of words and tokens sampled from the current dataset and persists when alternative density measures (frequency-weighted density, CELEX-based density) are used. Moreover, the effect is present in all 40 speakers in the Buckeye corpus, to more or less the same degrees, indicating that there is little individual difference. Overall these results suggest that the observed density effect is not likely to be artifactual of the current dataset or the specific density measure that is used. Instead, it represents a general trend that high-density words are shorter than low-density words, which can probably be replicated in other studies.

However, the duration study fails to find any reliable effect of average neighbor frequency, and the insignificance of this variable is confirmed by several model testing techniques, regardless of which dictionaries are used for coding neighbor frequency.

Vowel dispersion is measured as the Euclidean distance between the F1/F2 (in Bark) of the vowel token and that of the center of vowel space, normalized for vowel-specific de-

gree of dispersion. Center of vowel space is approximated by the average F1/F2 of four point vowels ([aa], [ae], [iy], [ow]). As shown in Chapter 5, the vowel dispersion model is overall less successful than the word duration model. Fixed-effects predictors including neighborhood measures make little contribution to the prediction of the outcome variable (i.e. K-dispersion), in terms of the overall proportion of variance that is explained by the model (R^2), though the proportion of variance that is attributed to random differences among words and speakers is significantly reduced after including the fixed-effects predictors. Put in another way, the inclusion of fixed-effects predictors does not improve the overall prediction accuracy of the model, but makes the prediction less random.

There are two main reasons for the reduced success of the vowel dispersion model. First of all, variation in vowel production is more complicated than durational variation. As mentioned in both Chapter 2 and the end of Chapter 5, vowel centralization/dispersion is not the *only* dimension that hypo-hyperarticulation in vowel production can occur. Speech reduction may also lead to vowel assimilation, resulting in vowel formants that are characteristic of the transition patterns to or from surrounding consonants, which are not necessarily closer to the center of vowel space. In that sense, even if certain linguistic factors are good predictors for hypo-hyperarticulation, a model on degree of vowel dispersion may not be able to reveal such relations. By contrast, durational variation can only occur along the shortening/lengthening dimension, and is therefore easier to model.

Secondly, the computation of K-dispersion is more complicated and therefore more prone to errors. Specifically, the K-dispersion measure is subject to alignment errors in the corpus transcription (for identifying vowel periods) and formant tracking errors in the LPC program. Moreover, the calculation of K-dispersion involves estimating the center of vowel space and approximating population means and standard deviations (for normalization). For comparison, the measurement of word duration is only subject to alignment errors in the transcription. As a result, the model on vowel dispersion may be less successful due to less accurate measurements.

Despite the overall reduced effectiveness, the vowel dispersion model does reveal some trends regarding neighborhood variables. The most robust finding is vowel centralization in words from dense neighborhoods. On average, words with many neighbors tend to have vowels that are closer to the center of vowel space than words with fewer neighbors. This effect is consistent with the shortening of high-density words, as found in the duration study. Model evaluation confirms that the observed effect is *not* confounded by phonotactic probability and it reliably exists in random subsets of the data. Moreover, the effect persists when neighborhood density is calculated from a different source (i.e. CELEX database)and exists in all but one speaker in the Buckeye corpus, although the degree of this effect varies considerably among speakers who are influenced.

Quite surprisingly, neighbor frequency is found with a positive effect on vowel dispersion in the baseline model. Other things being equal, words with high-frequency neighbors seem to have *more* dispersed vowels than words with low-frequency neighbors, and the size of this effect is similar to that of the density effect in the same model. However, when tested with CELEX-based neighborhood variables, neighbor frequency is no longer a significant

predictor for vowel dispersion (though neighborhood density remains significant, as stated above). Thus, the effect of neighbor frequency on vowel dispersion seems to be contingent on the coding of the variable. Given that the CELEX-based neighbor frequency represents a more up-to-date and sophisticated estimation, the fact that no effect is found with CELEX-based neighbor frequency seriously challenges the reliability of the neighbor frequency effect found in the baseline model.

To sum up, the current research has found that high neighborhood density is associated with shorter word duration and more centralized vowel production. By contrast, high neighbor frequency has no effect on word duration, and only a tendency for vowel dispersion, which is not stable across dictionaries.

6.2 Understanding the neighborhood effects

As presented in Chapter 2, the current research is set out to test two alternative hypotheses (repeated below in (1) and (2)) regarding the effects of neighborhood structure on word pronunciation.

(1) Speaker-oriented hypothesis

   (a) High-density words are easier to produce than low-density words. Words with high-frequency neighbors are easier to produce than those with low-frequency neighbors.

   (b) Easy-to-produce words tend to have shorter durations and more reduced vowels.

   (c) Therefore high-density words and words with high-frequency neighbors will be realized with *shorter* durations and *more reduced* vowels than low-density words and words with low-frequency neighbors, respectively.

(2) Listener-oriented hypothesis

   (a) High-density words are harder to recognize than low-density words. Words with high-frequency neighbors are harder to recognize than those with low-frequency neighbors.

   (b) Words with longer durations and more dispersed vowels are more intelligible.

   (c) Speakers monitor for listeners' needs and modify their speech accordingly.

   (d) Therefore high-density words and words with high-frequency neighbors will be realized with *longer* durations and *more dispersed* vowels than low-density words and words with low-frequency neighbors, respectively.

Most previous studies have supported some version of the listener-oriented hypothesis, based on the finding of vowel dispersion in words from dense neighborhoods. The current study also finds a tendency for words with high-frequency neighbors to have more expanded vowel space than words with low-frequency neighbors, but the overall range of findings is

much wider. As summarized in the preceding section, the current research investigates both durational variation and vowel dispersion, and both the effects of neighborhood density and the effects of neighbor frequency. Furthermore, the current research also tests with an alternative (i.e. CELEX-based) neighborhood measures, in addition to the widely-used HML metrics.

Variable	**HML**		**CELEX-based**	
	Word duration	**Vowel**	**Word duration**	**Vowel**
Neighborhood density	*Shortening*	*Centralizing*	*Shortening*	*Centralizing*
Neighbor frequency	*Null*	*Dispersing*	*Null*	*Null*

Table 6.1: Summary of the neighborhood effects found in the current work.

Thus, if we look at the complete map of current findings (see Table 6.1), it becomes clear that there is considerably more evidence for the speaker-oriented hypothesis than for the listener-oriented hypothesis. The most robust finding of the current research is the hypo-articulatory effects of neighborhood density, manifested in both durational shortening and vowel centralization, which agree with the predictions of the speaker-oriented hypothesis, but not the predictions of the listener-oriented hypothesis. If anything, shorter duration and more centralized vowel production are more likely to be detrimental, than conducive, to speech comprehension.

Thus, following the speaker-oriented hypothesis, words from dense neighborhoods are easier to produce, and therefore, tend to be reduced in speech. Recall from Chapter 2 that ease of production in high-density words can be ascribed to either easy lexical access (e.g. Vitevitch 2002; Vitevitch and Sommers 2003) or easy articulation due to high phonotactic probability (Munson 2001). In the current research, it has been repeatedly shown that the observed density effects are not likely to be fully confounded by phonotactic probability, which suggests that the locus of the effects cannot be only at the sublexical level (i.e. easy articulation). At least part of the neighborhood density effects must arise from the lexical level, due to facilitated lexical access in words from dense neighborhoods. Following this, one may ask *why would easy lexical access lead to phonetic reduction?* As discussed in Chapter 2, there is currently no ready answer to this question. A possible explanation is provided by Bell and colleagues (Bell et al. 2009), who hypothesized a general mechanism that keeps speech planning and articulation in synchrony. Under this account, words that are hard to access will be allocated with less time in articulation. However, whether or not this mechanism can be generalized to completely fluent speech is still in question.

6.3 Reconciling with previous findings

As mentioned before, previous studies (e.g. Kilanski 2009; Munson 2007; Munson and

Solomon 2004; Wright 1997, 2004) have all found vowel dispersion in words from dense neighborhoods, which is considered as evidence for listener adaptation. An immediate question is *how to understand the apparent discrepancy between previous and current findings?* To answer this question, it is important to point out a difference in speech style. While previous studies on neighborhood effects on vowel dispersion have all used speech data elicited in the lab, most typically using the word-reading paradigm, the current work uses speech data from spontaneously produced connected speech. It is a well-known fact that there are profound acoustic-phonetic differences between read speech and conversational speech. Words in conversational speech have shorter durations and more reduced vowels (e.g. Fosler-Lussier and Morgan 1999; Johnson 2004; Jurafsky et al. 1998; Yao et al. 2010), and are associated with lower intelligibility (Pickett and Pollack 1963). But the presence of context (both linguistic and nonlinguistic) may compensate for the reduced speech clarity and ensure successful communication. Another key difference between the two types of speech is that the speaker needs to allocate considerably more resources for planning the upcoming words in spontaneous speech than in read speech. As a consequence, fewer resources may be devoted to listener adaptation in spontaneous speech. Both the presence of context and the processing demand for speech planning may cause the speaker to be more inclined to a speaker-oriented approach than a listener-oriented approach in conversational speech production. That being said, it is not surprising that results from the current study, which uses spontaneous speech data, provide more support for the speaker-oriented hypothesis than for the listener-oriented hypothesis.

Furthermore, if we examine the experimental design in previous studies more closely, it becomes clear that previous and current findings are not completely incompatible. At least part of the apparent discrepancy can be attributed to differences in the dataset and the coding of neighborhood metrics. In the following, I will present several attempts to reconcile the current corpus results with previous experimental findings.

6.3.1 Reconciling with Wright (1997)

In Wright (1997) (as well as Experiment 1 of Munson and Solomon 2004), word stimuli were divided into "lexically hard" words, which have low frequency, high density and high neighbor frequency, and "lexically easy" words, which have high frequency, low density and low neighbor frequency. Given the current corpus results, we would expect high density to be associated with vowel centralization and high neighbor frequency (from HML) with vowel dispersion. Furthermore, the effects are probably of similar sizes so that they will tend to cancel each other out. Thus, the differences in vowel production between "lexically hard" and "lexically easy" words are mostly attributable to word frequency. Although word frequency is not a significant predictor in the vowel dispersion model of the current work[1], vowel reduction in high-frequency words is a robust finding that has been independently

[1]This is probably because the current study (both the dataset and the statistical model) is specifically designed for investigating neighborhood effects and therefore is not the most suitable for observing word frequency effects.

attested and replicated in the literature (Dinkin 2007; Fidelholtz 1975; Munson 2007, etc). Following this line, we would expect "lexically hard" words, which are lower in frequency, to have more dispersed vowels than "lexically easy" words, which have higher frequency. This coincides with what has been reported previously. Thus the results from Wright (1997, 2004) and Munson and Solomon (2004: Exp1) are in fact compatible with current findings.

6.3.2 Reconciling with Munson and Solomon (2004) and Munson (2007)

A different type of stimulus design was adopted by Experiment 2 of Munson and Solomon (2004) and Munson (2007). Instead of covarying all three lexical variables, word frequency was deliberately separated from neighborhood measures, but neighborhood density and neighbor frequency were still together, combined into a frequency-weighted density measure. Thus, the stimuli were divided into four conditions, by orthogonally combining high and low levels of frequency (HF/LF) and high and low levels of frequency-weighted density (HD/LD). Both experiments (which used the same set of stimuli) have found that for both HF and LF words, words from dense neighborhoods (HD) were produced with more dispersed vowels than those from sparse neighborhoods (LD).

However, if we calculate neighborhood density and neighbor frequency separately for the word stimuli (see Table 6.2), it seems that the differences between neighborhood density across conditions are quite small. Results from paired t-tests[2] show that the difference in neighborhood density between HD and LD words is only significant in low-frequency (LF) words ($t(19)$=2.83, p= 0.005), but not in high-frequency (HF) words ($t(19)$=0.84, p= 0.21). On the other hand, neighbor frequency is significantly different between HD and LD words in both LF ($t(19)$ = 3.53, p<0.001) and HF ($t(19)$=3.92, p<0.001) conditions. That is to say, the observed neighborhood effects are most reliably attributable to neighbor frequency, but not to neighborhood density. Thus, the reported findings in Munson and Solomon (2004) and Munson (2007) can be re-interpreted as vowel dispersion in words with high neighbor frequency (from HML), which is also found in the current research.

6.3.3 Reconciling with Kilanski (2009)

The recent dissertation work by Kilanski (2009) probably provides the greatest contrast to the current work. Neighborhood density is separated from both word frequency and neighbor frequency in the stimuli set. High-density word stimuli have on average 25.2 neighbors while low-density words have on average 13.8 neighbors, and the differences are statistically significant. Kilanski showed that words from high-density neighborhoods occupied a more expanded vowel space than words from low-density neighborhoods. Nevertheless, in the same

[2]Here I used paired t-tests because the word stimuli were matched in phonetic content across conditions. There were in total 20 quadruplets of CVC words. An example quadruplet is given in the following: *got* (HF/HD), *dock* (HF/LD), *dot* (LF/HD), *mop*(LF/LD)

Experimental condition	(raw) Neighborhood density	Neighbor frequency
HF/HD	23.50 (7.33)	2.17 (0.18)
HF/LD	21.80 (6.49)	1.96 (0.18)
LF/HD	24.15 (6.40)	2.10 (0.15)
LF/LD	18.9 (6.73)	1.88 (0.23)

Table 6.2: Mean values of raw neighborhood density and neighbor frequency of words in each experimental condition in the set of word stimuli used in Munson and Solomon (2004): Exp 2 and Munson (2007). Numbers in the parentheses are standard deviations.

work, Kilanski also presented other results which were obtained with highly similar experimental methods and the same subject population but seemed to link high neighborhood density with speech reduction. Most importantly, Kilanski found that whole word duration was *shorter* for high-density words than for low-density words - which agrees with the major finding from the word duration model of the current study.

To sum up, contrary to what seems on the surface, current findings from the Buckeye corpus do not necessarily contradict previous experimental results. As shown above, after separating the effects of neighborhood density and neighbor frequency (and word frequency as well), previous and current results may in fact be compatible, and the residual differences may be attributable to the differences between read speech and spontaneous speech.

6.4 Remaining puzzles

6.4.1 Differences between neighborhood density and neighbor frequency

One of the goals of the current research is to test the individual effects of neighborhood density and neighbor frequency on pronunciation variation. Results from the corpus studies do reveal significant differences between the two variables. Overall, neighborhood density has greater effects on pronunciation variation than neighbor frequency, which is not surprising given that density is more widely attested for its effects on word perception and production efficiency. What is not expected, however, is for high density and high neighbor frequency to show opposite tendencies regarding vowel dispersion. If the tendency regarding the hyperarticulatory effects of neighbor frequency proves to be tenable, the results would suggest that a densely populated neighborhood is *not* equivalent to a high-frequency neighborhood - while the former may lead to overall reduction in speech, the latter might lead to hyperarticulation, especially in vowel production. The exact source of this contrast is not clear at this moment. A speculative thought is that the contrast may have to do with the

mechanics of lexical activation and the spreading of activation during the process of speech production.

6.4.2 Relationship between duration and degree of vowel dispersion

Another remaining puzzle regards the relationship between variation in duration and variation in vowel dispersion. We know that shorter vowels tend to be more reduced than longer vowels, due to the duration-dependent vowel undershoot/overshoot (Moon and Lindblom 1994). But a less often asked question is what will the variation patterns be if this part of automatic correlation is controlled for. For example, if duration is statistically or empirically controlled, will vowels in high-frequency words still be more reduced, or less or equally reduced, compared with vowels in low-frequency words?

The current research has found that high neighborhood density leads to vowel centralization *beyond* durational shortening, and high neighbor frequency (from HML) seems cause vowel dispersion without affecting duration. A few other examples of this type of duration-*independent* vowel variation have been presented in previous research (Bell et al. 2003; Kilanski 2009; Krause and Braida 2002, 2004), but generally speaking, little is known about such variation, not to mention what features of the production system may be responsible for it. This can be a potential problem for the speaker-oriented account of pronunciation variation[3], because no hypothesis can be formed beyond duration-dependent phonetic variation.

6.5 Conclusion

To conclude, this dissertation contains two corpus studies on the effects of neighborhood structure on pronunciation variation. The most robust findings are word shortening and vowel centralization in words from dense neighborhoods. These findings provide strong evidence for the speaker-oriented hypothesis of pronunciation variation over the listener-oriented hypothesis. The source of the density effects is probably in lexical access. Words from dense neighborhoods receive more activation from neighbors than words from sparse neighborhoods. As a result, high-density words are easier to access than low-density words, and the ease of lexical access may lead to phonetic reduction, in order to keep a synchrony between speech planning and execution.

The current research also reveals a peripheral finding, that is, vowels in words with high-frequency neighbors tend to have more dispersed vowels. This tendency is consistent with previous experimental results, but the current research shows that it is only significant when a particular set of neighborhood measures (from the Hoosier mental lexicon) is used. Thus, the reliability of this effect remains to be tested in future research.

Overall, the current findings suggest that speaker-oriented forces have greater impacts on word-level pronunciation variation, compared with listener-oriented forces. Everything

[3]Crucially, even the proposal in Bell et al. (2009) for linking ease of planning with rate of speech production is based on temporal coordination.

else being equal, the pronunciation of a word is more likely to be influenced by features of the production system, which are in general not under the voluntary control of the speaker, than by the consideration of listeners' needs. Speech adaptation to the listener may occur on a global level (e.g. child-oriented and foreigner-oriented speech, and speech under noise), but in terms of word-by-word pronunciation variation, listener orientation tends to be overridden by the constraints of the speaker's own production system. That being said, however much we would like to be truly altruistic human beings, we are just human beings after all.

Part I

References

References

Alario, F., L. Perre, C. Castel, and J. Ziegler (2007). The role of orthography in speech production revisited. *Cognition 102*(3), 464–475.

Allopenna, P., J. Magnuson, and M. Tanenhaus (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language 38*(4), 419–439.

Almeida, J., M. Knobel, M. Finkbeiner, and A. Caramazza (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. *Psychonomic Bulletin & Review 14*(6), 1177–1182.

Arbesman, S., S. Strogatz, and M. Vitevitch (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos 20*(3), 679–685.

Aylett, M. and A. Turk (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech 47*(1), 31–56.

Aylett, M. and A. Turk (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of Acoustical Society of America 119*(5), 3048–3058.

Baars, B. and M. Motley (1974). Spoonerisms: Experimental elicitation of human speech errors. *Catalog of Selected Documents in Psychology 4*, 118. Abstract obtained from Journal Supplement Abstract Service.

Baars, B. J. (1992). A dozen competing-plans techniques for inducing predictable slips in speech and action. In B. J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition*, pp. 129–150. New York: Plenum Press.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2009). *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. R package version 0.955. http://CRAN.R-project.org/package=languageR.

Baayen, R. H., R. Piepenbrock, and H. V. Rijn (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Baese-Berk, M. and M. Goldrick (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes 24*(4), 527 – 554.

Baldwin, A. and C. Baldwin (1973). The study of mother-child interaction. *American Scientist 61*(6), 714–721.

Balota, D., J. Boland, and L. Shields (1989). Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language 28*(1), 14–36.

Balota, D. and J. Chumbley (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language 24*(1), 89–106.

Bard, E. C. and M. P. Aylett (2005). Referential form, word duration, and modeling the listener in spoken dialogue. In J. C. Trueswell and M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pp. 173 – 191. Cambridge, Massachusetts: MIT Press.

Bard, E. G., A. H. Anderson, M. Aylett, G. Doherty-Sneddon, and A. Newlands (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language 42*(1), 1–22.

Bard, E. G. and R. C. Shillcock (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. Altmann and R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, pp. 235 – 275. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Barry, C., K. Hirsh, R. Johnston, and C. Williams (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of memory and language 44*(3), 350–375.

Bartram, D. (1973). The effects of familiarity and practice on naming pictures of objects. *Memory & Cognition 1*, 101–105.

Bartram, D. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology 6*(3), 325–356.

Bates, D. and M. Maechler (2010). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-33. http://CRAN.R-project.org/package=lme4.

Baus, C., A. Costa, and M. Carreiras (2008). Neighbourhood density and frequency effects in speech production: A case for interactivity. *Language and Cognitive Processes 23*(6), 866–888.

Beckman, M. (1986). *Stress and non-stress accent*. Dordrecht: Fortis.

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

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand, and D. Gildea (1999). Forms of English function words - Effects of disfluencies, turn position, age and sex, and predictability. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS XIV)*, San Francisco, USA, pp. 395–398.

Bell, A., D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America 113*(2), 1001–1024.

Biggs, T. and H. Marmurek (1990). Picture and word naming: Is facilitation due to processing overlap? *The American Journal of Psychology 103*(1), 81–100.

Boersma, P. and D. Weenink (2008). *Praat: Doing phonetics by computer*. Version 5.0.26. http://www.praat.org.

Bond, Z. S. and T. J. Moore (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication 14*(4), 325–337.

Bonin, P. and M. Fayol (2002). Frequency effects in the written and spoken production of homophonic picture names. *European Journal of Cognitive Psychology 14*(3), 289–313.

Boothroyd, A. and S. Nittrouer (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America 84*(1), 101–114.

Bowdle, B. F. and R. Wright (1998). Lexical neighborhoods and subjective intelligibility ratings: A preliminary report. In *Research on Speech Perception Progress Report No. 22*, pp. 345–352. Bloomington: Speech Research Laboratory, Psychology Department, Indiana University.

Bradlow, A. and T. Bent (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America 112*(1), 272–284.

Bradlow, A. and D. B. Pisoni (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of Acoustical Society of America 106*(4), 2074–2085.

Bradlow, A. R., G. Torretta, and D. Pisoni (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication 20*(3–4), 255–272.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts (Eds.), *Cognitive foundations of interpretation*, pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and M. Ford (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language 86*(1), 168–213.

Broen, P. (1972). The verbal environment of the language-learning child. *ASHA Monography 17*.

Browman, C. and L. Goldstein (1992). Articulatory phonology: An overview. *Phonetica 49*(3-4), 155–180.

Brown, A. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin 109*(2), 204–223.

Brown, R. and D. McNeill (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior 5*(4), 325–337.

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

Bybee, J. (2000). The phonology of the lexicon: Evidence from lexical diffusion. In M. Barlow and S. Kemmer (Eds.), *Usage-based models of language*, pp. 65–85. Stanford: CSLI.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2002a). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition 24*(2), 215–222.

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

Byrd, D., A. Kaun, S. Narayanan, and E. Saltzman (2000). Phrasal signatures in articulation. In *Papers in Laboratory Phonology V*, pp. 70–87. Cambridge: Cambridge University Press.

Carroll, J. and M. White (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology 25*(1), 85–95.

Chan, K. and M. Vitevitch (2009). The influence of the phonological neighborhood clustering-coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance 35*(6), 1934–1949.

Chan, K. and M. Vitevitch (2010). Network structure influences speech production. *Cognitive Science 34*(4), 685–697.

Charles-Luce, J. and P. A. Luce (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language 17*(1), 205– 215.

Clark, H. and J. Fox Tree (2002). Using uh and um in spontaneous speaking. *Cognition 84*(1), 73–111.

Clark, H. and C. Marshall (1981). Definite reference and mutual knowledge. In A. Joshe, B. Webber, and I. Sag (Eds.), *Elements of Discourse Understanding*, pp. 10–63. Cambridge: Cambridge University Press.

Clifton, C. et al. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior 23*(6), 696–708.

Cluff, M. and P. Luce (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance 16*(3), 551–563.

Coady, J. and R. Aslin (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language 30*(2), 441–469.

Cole, R. and C. Perfetti (1980). Listening for mispronunciations in a children's story: The use of context by children and adults. *Journal of Verbal Learning and Verbal Behavior 19*(3), 297–315.

Colombo, L., M. Pasini, and D. A. Balota (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition 34*(6), 1312 – 1324.

Coltheart, M., E. Davelaar, J. Jonasson, and D. Besner (1976). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*, pp. 535–555. Hillsdale, New Jersey: Erlbaum.

Costa, A. and N. Sebastian-Galles (1998). Abstract phonological structure in language production: Evidence from Spanish. *Journal of Experimental Psychology: Learning, Memory, and Cognition 24*(4), 886–903.

Craig, C., B. Kim, P. Rhyner, and T. Chirillo (1993). Effects of word predictability, child development, and aging on time-gated speech recognition performance. *Journal of Speech and Hearing Research 36*(4), 832.

Crystal, T. and A. House (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America 88*(1), 101–112.

Cutler, A. and S. Butterfield (1991). Word boundary cues in clear speech: A supplementary report. *Speech Communication 10*(4), 335–353.

Cutler, A. and D. Norris (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance 14*(1), 113–121.

Dahan, D., J. S. Magnuson, and M. K. Tanenhaus (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology 42*(4), 317 – 367.

Damian, M. and J. Bowers (2003). Effects of orthography on speech production in a form-preparation paradigm. *Journal of Memory and Language 49*(1), 119–132.

De Cara, B. and U. Goswami (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods Instruments and Computers 34*(3), 416 – 423.

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

Dell, G. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes 5*(4), 313–349.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review 93*(3), 283–321.

Dell, G. S. and J. K. Gordon (2003). Neighbors in the lexicon: Friends or foes. In N. O. Schiller and A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities*, Volume 6, pp. 9 – 37. Berlin, New York: Mouton de Gruyter.

Dinkin, A. (2007). The real effect of word frequency on phonetic variation. In *Proceedings of the 31st Penn Linguistics Colloquium*, Philadelphia, Pennsylvania.

Dirks, D. D., S. Takayanagi, A. Moshfegh, P. D. Noffsinger, and S. A. Fausti (2001). Examination of the neighborhood activation theory in normal and hearing-impaired listeners. *Ear and Hearing 22*(1), 1–13.

Duez, D. (1993). Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research 22*(1), 21–39.

Durso, F. and M. Johnson (1979). Facilitation in naming and categorizing repeated pictures and words. *Journal of Experimental Psychology: Human Learning and Memory 5*(5), 449–459.

Eklund, R. and E. Shriberg (1998). Crosslinguistic disfluency modelling: A comparative analysis of Swedish and American English human–human and human–machine dialogues. In *Proceedings of the International Conference on Spoken Language Processing*, Volume 6, pp. 2631–2634. Australian Speech Science and Technology Association.

Ernestus, M., M. Lahey, F. Verhees, and R. Baayen (2006). Lexical frequency and voice assimilation. *The Journal of the Acoustical Society of America 120*(2), 1040–1051.

Fay, D. and A. Cutler (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry 8*(3), 505–520.

Ferrand, L. (1996). The masked repetition priming effect dissipates when increasing the inter-stimulus interval: Evidence from word naming. *Acta Psychologica 91*(1), 15–25.

Ferreira, F. and K. Bailey (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences 8*(5), 231–237.

Ferreira, V. and G. Dell (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology 40*(4), 296–340.

Fidelholtz, J. (1975). Word frequency and vowel reduction in English. In *Papers from the 13th Regional Meeting, Chicago Linguistic Society*, pp. 200–213.

Forster, K. and S. Chambers (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior 12*(6), 627–635.

Forster, K. and C. Davis (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition 10*(4), 680–698.

Fosler-Lussier, E. and N. Morgan (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication 29*(2), 137–158.

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

Fowler, C. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech 31*(4), 307.

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

Fox Tree, J. and H. Clark (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition 62*(2), 151–167.

Francis, W., B. Augustini, and S. Sáenz (2003). Repetition priming in picture naming and translation depends on shared processes and their difficulty: Evidence from spanish–english bilinguals. *Learning, Memory 29*(6), 1283–1297.

Francis, W. and S. Sáenz (2007). Repetition priming endurance in picture naming and translation: Contributions of component processes. *Memory & Cognition 35*(3), 481.

Frank, A. and T. Jaeger (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual conference of the cognitive science society*, pp. 933–938.

Frauenfelder, U. H., R. H. Baayen, F. M. Hellwig, and R. Schreuder (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language 32*(6), 781 – 804.

Frisch, S., N. Large, and D. Pisoni (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language 42*(4), 481–496.

Gahl, S. (2008). *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language 84*(3), 474–496.

Gahl, S. and S. Garnsey (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language 80*(4), 748–775.

Garlock, V., A. Walley, and J. Metsala (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language 45*(3), 468–492.

Gaskell, M. and W. Marslen-Wilson (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes 12*(5), 613–656.

Goldinger, S., P. Luce, and D. Pisoni (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language 28*(5), 501–518.

Goldinger, S., P. Luce, D. Pisoni, and J. Marcario (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition 18*(6), 1211–1211.

Goldinger, S. and W. V. Summers (1989). Lexical neighborhoods in speech production: A first report. *The Journal of the Acoustical Society of America 85*, S97.

Goldwater, S., D. Jurafsky, and C. Manning (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication 52*(3), 181–200.

Greenberg, J. and J. Jenkins (1967). Studies in the psychological correlates of the sound system of American English. In L. A. Jacobovits and M. S. Miron (Eds.), *Readings in the psychology of language*, pp. 186–200. New Jersey: Prentice Hall.

Greenberg, S., H. Carvey, and L. Hitchcock (2002). The relation between stress accent and pronunciation variation in spontaneous american english discourse. In *Proceedings of the ISCA workshop on prosody and speech processing*, Aix-en-Provence, France.

Gregory, M. L., W. D. Raymond, A. Bell, E. Fosler-Lussier, and D. Jurafsky (1999). The effects of collocational strength and contextual predictability in lexical production. In *Proceedings of the Chicago Linguistic Society 35*, pp. 151 – 166.

Griffin, Z. (2002). Recency effects for meaning and form in word selection. *Brain and Language 80*(3), 465–487.

Griffin, Z. and K. Bock (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language 38*(3), 313–338.

Grosjean, F. and J. Fitzler (1984). Can semantic constraint reduce the role of word frequency during spoken-word recognition? *Bulletin of the Psychonomic Society 22*(3), 180–182.

Hagiwara, R. (1997). Dialect variation and formant frequency: The american english vowels revisited. *The Journal of the Acoustical Society of America 102*(1), 655–658.

Harley, T. and H. Bown (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology 89*(1), 151–174.

Hartley, H. and J. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika 54*, 93–108.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics 31*(3-4), 373–405.

Hillenbrand, J., L. A. Getty, M. J. Clark, and K. Wheeler (1995). Acoustic characteristics of american english vowels. *Journal of the Acoustical Society of America 97*(5), 3099 – 3111.

Hollich, G., P. Jusczyk, and P. Luce (2002). Lexical neighborhood effects in 17-month-old word learning. In *Proceedings of the 26th Annual Boston University Conference on Language Development*, Boston, MA, pp. 314–323. Cascadilla Press.

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (Ed.), *Current progress in historical linguistics*, pp. 96–105. Amsterdam: North Holland.

Horton, W. and B. Keysar (1996). When do speakers take into account common ground? *Cognition 59*(1), 91–117.

Howard, S. and J. Burt (2010). Evidence from the attentional blink for different sources of word repetition effects. *Consciousness and Cognition 19*(1), 125–134.

Hunnicut, S. (1985). Intelligibility versus redundancy–conditions of dependency. *Language and Speech 28*(1), 47–56.

James, L. and D. Burke (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition 26*(6), 1378–1391.

Jescheniak, J. and W. Levelt (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition 20*(4), 824–843.

Jescheniak, J. and H. Schriefers (2001). Priming effects from phonologically related distractors in picture–word interference. *The Quarterly Journal of Experimental Psychology Section A 54*(2), 371–382.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama and K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th International Symposium*, Tokyo, Japan, pp. 29–54. The National International Institute for Japanese Language.

Jones, G. and S. Langford (1987). Phonological blocking in the tip of the tongue state. *Cognition 26*(2), 115–122.

Junqua, J. (1996). The influence of acoustics on speech production: a noise-induced stress phenomenon known as the lombard reflex. *Speech Communication 20*(1-2), 13–22.

Jurafsky, D., A. Bell, E. Fosler-Lussier, C. Girand, and W. Raymond (1998). Reduction of English function words in Switchboard. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98)*, Volume 7, Sydney, Australia, pp. 3111–3114.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond (2001a). The effect of language model probability on pronunciation reduction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Volume 2, Salt Lake City, Utah, pp. 801–804.

Jurafsky, D., A. Bell, M. Gregory, and W. Raymond (2001b). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure*, pp. 229–254. Amsterdam: John Benjamins.

Jusczyk, P. and P. Luce (2002). Speech perception and spoken word recognition: Past and present. *Ear and Hearing 23*(1), 2–40.

Jusczyk, P. W., P. A. Luce, and J. Charles-Luce (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language 33*(5), 630–645.

Kiesling, S., L. Dilley, and W. D. Raymond (2006). *The Variation in Conversation (ViC) project: Creation of the Buckeye corpus of conversational speech.* Department of Psychology, Ohio State University. http://www. buckeyecorpus.osu.edu.

Kilanski, K. (2009). *The effects of token frequency and phonological neighborhood density on native and non-native English speech production.* Ph. D. thesis, University of Washington.

Kirby, J. and A. Yu (2007). Lexical and phonotactic effects on wordlikeness judgements in Cantonese. In *Proceedings of the 16th International Congress of the Phonetic Sciences (ICPhS XVI)*, Saarbrücken, Germany, pp. 1161–1164.

Klatt, D. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics 3*(3), 129–140.

Klatt, D. (1979). Synthesis by rule of segmental durations in english sentences. In B. Lindblom and S. Ohman (Eds.), *Frontiers of speech communication research*, pp. 287–299. New York: Academic Press.

Krause, J. C. and L. D. Braida (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America 112*(5), 2165 – 2172.

Krause, J. C. and L. D. Braida (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America 115*(1), 362 – 378.

Kučera, H. and W. Francis (1967). *Computational analysis of present-day American English.* Providence, Rhode Island: Brown University Press.

La Heij, W., M. Puerta-Melguizo, M. van Oostrum, and P. Starreveld (1999). Picture naming: Identical priming and word frequency interact. *Acta Psychologica 102*(1), 77–95.

Labov, W., S. Ash, and C. Boberg (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool.* Walter De Gruyter Inc.

Ladd, D. and N. Campbell (1991). Theories of prosodic structure: evidence from syllable duration. In *Proceedings of the 12th International Congress of Phonetic Sciences (ICPhS XII)*, Volume 2, Aix-en-Provence, France, pp. 290–293.

Ladefoged, P. (1967). *Three areas of experimental phonetics.* London: Oxford University Press.

Ladefoged, P. (2001). *A course in phonetics [fourth edition].* Heinle & Heinle, Thompson Learning.

Landauer, T. and L. Streeter (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior 12*(2), 119–131.

Lau, P. (2008). The lombard effect as a communicative phenomenon. In *UC Berkeley Phonology Lab Report*, pp. 1–9. Phonology Lab, Linguistics Department, University of California at Berkeley.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Massachusetts: MIT Press.

Leslau, W. (1969). Frequency as determinant of linguistic changes in the Ethiopian languages. *Word 25*, 180–189.

Levelt, W., P. Praamstra, A. Meyer, P. Helenius, and R. Salmelin (1998). An meg study of picture naming. *Journal of Cognitive Neuroscience 10*(5), 553–567.

Levelt, W., A. Roelofs, and A. Meyer (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences 22*(1), 1–38.

Levy, R. and T. Jaeger (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, and T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS)*, Volume 19, pp. 849–856. Cambridge, Massachusetts: MIT Press.

Lickley, R. (1994). *Detecting disfluency in spontaneous speech*. Ph. D. thesis, University of Edinburgh.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech 6*(3), 172–187.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America 35*(5), 783.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. J. Hardcastle and A. Marchal (Eds.), *Speech production and speech modelling*, pp. 403–439. Dordrecht, The Netherlands: Kluwer.

Lively, S. E., D. B. Pisoni, W. V. Summers, and R. H. Bernacki (1993). Effects of cognitive workload on speech production. *Journal of Acoustical Society of America 93*(5), 2962–2973.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Annales de Maladies de L'oreille et du Larynx 37*, 101–119.

Long, M. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences 379 (Native Language and Foreign Language Acquisition)*, 259–278.

Luce, P. (1986). Neighborhoods of words in the mental lexicon. In *Research on Speech Perception Technical Report No. 6*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Luce, P., S. Goldinger, E. Auer, and M. Vitevitch (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics 62*(3), 615–625.

Luce, P. and N. Large (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes 16*(5), 565–581.

Luce, P. A. and D. B. Pisoni (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing 19*(1), 1–36.

Lupker, S. J. (1982). The role of phonetic and orthographic similarity in picture-word interference. *Canadian Journal of Psychology 36*(3), 349–376.

MacKay, D. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review 89*(5), 483–506.

MacKay, D. (1988). The organization of perception and action. a theory for language and other cognitive skills. *The Italian Journal of Neurological Sciences 9*(3), 303–303.

Marian, V. and H. Blumenfeld (2006). Phonological neighborhood density guides: Lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries 2*, 3–35.

Marian, V., H. Blumenfeld, and O. Boukrina (2008). Sensitivity to phonological similarity within and across languages. *Journal of Psycholinguistic Research 37*(3), 141–170.

Marslen-Wilson, W. (1985). Speech shadowing and speech comprehension. *Speech Communication 4*(1-3), 55–73.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition 25*(1-2), 71–102.

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In G. Altmann (Ed.), *Cognitive models of speech processing*, Cambridge, Massachusetts, pp. 148–172. MIT Press.

Marslen-Wilson, W. and L. Tyler (1980). The temporal structure of spoken language understanding. *Cognition 8*(1), 1–71.

Marslen-Wilson, W. and A. Welsh (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology 10*(1), 29–63.

McClelland, J. and J. Elman (1986). The trace model of speech perception. *Cognitive Psychology 18*(1), 1–86.

McClelland, J. and D. Rumelhart (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review 88*(5), 375–407.

Meyer, A. (1996). Lexical access in phrase and sentence production: Results from picture-word interference experiments. *Journal of Memory and Language 35*(4), 477–496.

Meyer, A. and K. Bock (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation? *Memory & Cognition 20*(6), 715–715.

Monsell, S., G. Matthews, and D. Miller (1992). Repetition of lexicalization across languages: A further test of the locus of priming. *The Quarterly Journal of Experimental Psychology Section A 44*(4), 763–783.

Moon, S. and B. Lindblom (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of Acoustical Society of America 96*(1), 40–55.

Moss, H., R. Ostrin, L. Tyler, and W. Marslen-Wilson (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning Memory and Cognition 21*(4), 863–883.

Motley, M. and B. Baars (1976). Laboratory induction of verbal slips: A new method for psycholinguistic research. *Communication Quarterly 24*(2), 28–34.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research 44*(4), 778–792.

Munson, B. (2007). Lexical access, lexical representation, and vowel production. In J. Cole and J. Hualde (Eds.), *Papers in Laboratory Phonology VIIII*. Walter De Gruyter.

Munson, B. and N. P. Solomon (2004). The influence of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research 47*(2), 1048 – 1058.

Myers, J. and Y. Li (2009). Lexical frequency effects in Taiwan Southern Min syllable contraction. *Journal of Phonetics 37*(2), 212–230.

Neu, H. (1980). Ranking of constraints on/t, d/deletion in American English: A statistical analysis. In W. Labov (Ed.), *Locating language in time and space*, pp. 37–54. New York: Academic.

New, B., L. Ferrand, C. Pallier, and M. Brysbaert (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review 13*(1), 45.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition 52*(3), 189–234.

Nusbaum, H. C., D. B. Pisoni, and C. K. Davis (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10*. Bloomington, Indiana: Speech Reserach Laboratory, Psychology Department, Indiana University.

Oldfield, R. C. and A. Wingfield (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology 17*, 273–281.

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, Volume 1, San Francisco, USA, pp. 593–596.

Patterson, D. and C. Connine (2000). Variant frequency in flap production. *Phonetica 58*(4), 254–275.

Payton, K. L., R. M. Uchanski, and L. D. Braida (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of Acousitical Society of America 95*(3), 1581–1592.

Perfect, T. and J. Hanley (1992). The tip-of-the-tongue phenomenon: Do experimenter-presented interlopers have any effect? *Cognition 45*(1), 55–75.

Peterson, G. and H. Barney (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America 24*(2), 175–184.

Phillips, B. (1980). Old English an on: A new appraisal. *Journal of English Linguistics 14*(1), 20–23.

Phillips, B. (1981). Lexical diffusion and Southern *tune, duke, news*. *American Speech 56*(1), 72–78.

Phillips, B. (1983). ME diphthongization, phonetic analogy, and lexical diffusion. *Word 34*, 11–24.

Phillips, B. (1984). Word frequency and the actuation of sound change. *Language 60*(2), 320–342.

Picheny, M. A., N. I. Durlach, and L. D. Braida (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research 28*(1), 96–103.

Picheny, M. A., N. I. Durlach, and L. D. Braida (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.

Pickett, J. and I. Pollack (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech 6*(3), 151–164.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven, T. Rietvelt, and N. Warner (Eds.), *Papers in Laboratory Phonology VII*, pp. 101–139. Berlin, New York: Mouton de Gruyter.

Pisoni, D., H. Nusbaum, P. Luce, and L. Slowiaczek (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication 4*(1–3), 75–95.

Pitt, M. A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Department of Psychology, Ohio State University. http://www.buckeyecorpus.osu.edu.

Plaut, D., J. McClelland, M. Seidenberg, and K. Patterson (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review 103*(1), 56–115.

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

Radeau, M., J. Morais, and A. Dewier (1989). Phonological priming in spoken word recognition: task effects. *Memory & Cognition 17*(5), 525–535.

Raymond, W. D., R. Dautricourt, and E. Hume (2006). Modeling the effects of extralinguistic, lexical, and phonological factors. *Language Variation and Change 18*(1), 55–97.

Remick, H. (1971). *The maternal environment of language acquisition*. Ph. D. thesis, University of California at Davis.

Rhodes, R. (1992). Flapping in American English. In W. U. Dressler, M. Prinzhorn, and J. Rennison (Eds.), *Proceedings of the 7th International Phonology Meeting*, Torino, pp. 217–232. Rosenberg & Sellier.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition 42*(1-3), 107–142.

Roelofs, A. (2006). The influence of spelling on phonological encoding in word reading, object naming, and word generation. *Psychonomic Bulletin & Review 13*(1), 33–37.

Sachs, J., R. Brown, and R. Salerno (1976). Adults' speech to children. In W. van Raffler Engel and Y. LeBrun (Eds.), *Baby talk and infant speech (Neurolinguistics 5)*, pp. 240–245. Lisse, Netherlands: Swetz & Zeitlinger.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. R package version 0.18-8. http://lmdvr.r-forge.r-project.org.

Scarborough, D., C. Cortese, and H. Scarborough (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance 3*(1), 1–17.

Scarborough, R. (2004). Degree of coarticulation and lexical confusability. In *Proceedings of the 29th Meeting of the Berkeley Linguistics Society*, Berkeley.

Scarborough, R. A. (in press). Lexical and contextual predictability: Confluent effects on the production of vowels. In *Papers in Laboratory Phonology X: Variation, Phonetic Detail, and Phonological Modeling*. Mouton de Gruyter.

Schriefers, H., A. Meyer, and W. Levelt (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language 29*(1), 86–102.

Schuchardt, H. (1885). *Über die Lautgesetze: gegen die Junggrammatiker*. Berlin: R. Oppenheim.

Sereno, J. and A. Jongman (1990). Phonological and form class relations in the lexicon. *Journal of Psycholinguistic Research 19*(6), 387–404.

Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition 42*(1-3), 213–259.

Shields, L. and D. Balota (1991). Repetition and associative context effects in speech production. *Language and Speech 34*(1), 47.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS XIV)*, Volume 1, San Francisco, USA, pp. 619–622.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association 31*(1), 153 – 169.

Silipo, R. and S. Greenberg (1999). Automatic transcription of prosodic stress for spontaneous english discourse. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS XIV)*, San Francisco, USA, pp. 2351–2354.

Sinclair, J. (Ed.) (1987). *Looking Up: An account of the COBUILD Project in lexical computing*. London: Collins.

Slowiaczek, L., H. Nusbaum, and D. Pisoni (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition 13*(1), 64–75.

Slowiaczek, L. and D. Pisoni (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory & Cognition 14*(3), 230–237.

Snow, C. and C. Ferguson (Eds.) (1977). *Talking to children: Language input and acquisition*. Cambridge: Cambridge University Press.

Sommers, M. and S. Danielson (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging 14*(3), 458–472.

Sommers, M. S., K. I. Kirk, and D. B. Pisoni (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. i: The effects of response format. *Ear and Hearing 18*(2), 89–99.

Stemberger, J. (1985). An interactive activation model of language production. In A. Ellis (Ed.), *Progress in the psychology of language*, Volume 1, pp. 143–186. London: Erlbaum.

Stemberger, J. (1991). Apparent anti-frequency effects in language production: The addition bias and phonological underspecification. *Journal of Memory and Language 30*(2), 161–185.

Stemberger, J. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language 90*(1-3), 413–422.

Stemberger, J. and R. Treiman (1986). The internal structure of word-initial consonant clusters. *Journal of Memory and Language 25*(2), 163–180.

Storkel, H. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language 29*(02), 251–274.

Storkel, H. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics 25*(02), 201–221.

Storkel, H. (2006). Do children still pick and choose? The relationship between phonological knowledge and lexical acquisition beyond 50 words. *Clinical Linguistics & Phonetics 20*(7-8), 523–529.

Summers, W., D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America 84*(3), 917–928.

Tanenhaus, M., H. Flanigan, and M. Seidenberg (1980). Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition 8*(6), 513–520.

Tanenhaus, M., L. Stowe, and G. Carlson (1985). The interaction of lexical expectation and pragmatics in parsing filler-gap constructions. In *Proceedings of the 7th Annual Cognitive Science Society Meetings*, Hillsdale, New Jersey, pp. 361–365. Erlbaum.

Tily, H., S. Gahl, I. Arnon, N. Snider, A. Kothari, and J. Bresnan (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition 1*(2), 147–165.

Trueswell, J., M. Tanenhaus, and C. Kello (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition 19*(3), 528–528.

Tsai, P. T. (2007). The effects of phonological neighborhoods on spoken word recognition in Mandarin Chinese. Master's thesis, University of Maryland.

Uchanski, R. M., S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach (1996). Speaking clearly for the hard of hearing IV: further studies of the role of speaking rate. *Journal of Speech and Hearing Research 39*(3), 494–509.

Van Bergem, D. (1995). Acoustic and lexical vowel reduction. In *Studies in Language and Language Use*, Volume 16. Amsterdam: IFOTT.

Van Son, R. and L. Pols (2003). How efficient is speech. In *Proceedings of the Institute of Phonetic Sciences*, Volume 25, pp. 171–184.

Ventura, P., R. Kolinsky, J. Querido, S. Fernandes, and J. Morais (2007). Is phonological encoding in naming influenced by literacy? *Journal of Psycholinguistic Research 36*(5), 341–360.

Vitevitch, M. (1997). The neighborhood characteristics of malapropisms. *Language and Speech 40*(3), 211.

Vitevitch, M. (2007). The spread of the phonological neighborhood influences spoken word recognition. *Memory & Cognition 35*(1), 166–175.

Vitevitch, M. and P. Luce (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science 9*(4), 325–327.

Vitevitch, M. and P. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language 40*(3), 374–408.

Vitevitch, M. and M. Stamer (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes 21*(6), 760–770.

Vitevitch, M. and M. Stamer (2009). The influence of neighborhood density (and neighborhood frequency) in spanish speech production: A follow-up report. In *Spoken, Language Laboratory Technical Report*, Volume 1, pp. 1–6. University of Kansas.

Vitevitch, M., M. Stamer, and J. Sereno (2008). Word length and lexical competition: Longer is the same as shorter. *Language and Speech 51*(4), 361.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning Memory and Cognition 28*(4), 735–747.

Vitevitch, M. S., J. Ambrüster, and S. Chu (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition 30*(2), 514–529.

Vitevitch, M. S. and P. A. Luce (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers 36*(3), 481–487.

Vitevitch, M. S., P. A. Luce, D. B. Pisoni, and E. T. Auer (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language 68*, 306–311.

Vitevitch, M. S. and M. S. Sommers (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition 31*(4), 491–504.

Warner, N., A. Jongman, J. Sereno, and R. Kemps (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics 32*(2), 251 – 276.

Warren, P. and W. Marslen-Wilson (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics 41*(3), 262–275.

Watson, D., M. Breen, and E. Gibson (2006). The role of syntactic obligatoriness in the production of intonational boundaries. *Journal of Experimental Psychology: Learning, Memory, and Cognition 32*(5), 1045 – 1056.

Watson, P. and B. Munson (2007). A comparison of vowel acoustics between older and younger adults. In *Proceedings of the 16th International Congress of the Phonetic Sciences (ICPhS XVI)*, Saarbrücken, Germany.

Weber, A. and A. Cutler (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language 50*(1), 1–25.

Wheeldon, L. and S. Monsell (1992). The locus of repetition priming of spoken word production. *The Quarterly Journal of Experimental Psychology Section A 44*(4), 723–761.

Wingfield, A. (1967). Perceptual and response hierarchies in object identification. *Acta Psychologica 26*, 216–226.

Woodley, M. (2010). For preschoolers, lexical access is purely lexical: Neighborhood density effects in child speech perception and the emergent phoneme hypothesis. In *UC Berkeley Phonology Lab Report*, Number 395–405. Phonology Lab, Linguistics Department, University of California at Berkeley.

Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. In *Research on Speech Perception Progress Report No. 21*, pp. 471–485. Bloomington, Indiana: Speech Research Laboratory, Psychology Department, Indiana University.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, and R. Temple (Eds.), *Papers in Laboratory Phonology VI*, pp. 26–50. Cambridge: Cambridge University Press.

Yao, Y., S. Tilsen, R. L. Sprouse, and K. Johnson (2010). Automated measurement of vowel formants in the buckeye corpus. *Gengo Kenkyu 138*, 99–113.

Yoneyama, K. (2002). *Phonological neighborhoods and phonetic similarity in Japanese word recognition*. Ph. D. thesis, Ohio State University.

Yoneyama, K. and B. Munson (2010). The influence of lexical factors on word recognition by native english speakers and japanese speakers acquiring english: A first report. *Journal of the Phonetic Society of Japan 14*, 35–47.

Ziegler, J. and L. Ferrand (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review 5*(4), 683–689.

Ziegler, J., L. Ferrand, and M. Montant (2004). Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory & Cognition 32*(5), 732–741.

Zipf, G. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 40*, 1–95.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Acoustical Society of America Journal 33*(2), 248.

Zwicker, E. (1975). Scaling. In W. D. Keidel and W. D.Neff (Eds.), *Auditory System: Physiology (CNS), behavioral studies, psychoacoustics*. Berlin: Springer-Verlag.

Zwicker, E. and E. Terhardt (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America 68*(5), 1523.

Zwitserlood, C. (1989). *Words and sentences: The effects of sentential-semantic context on spoken-word processing*. Ph. D. thesis, Katholieke Universiteit Nijmegen.

Part II

Appendices

Appendix A

Word lists

A.1 Target word list (n=540)

back, bad, badge, bag, ball, bar, bare, base, bash, bass, bat, bath, beach, bean, bear, beat, bed, beer, bell, berth, big, bike, bill, birth, bitch, bite, boat, bob, boil, bomb, book, boom, boon, boss, bought, bout, bowl, buck, bull, bum, burn, bus, bush, cab, cad, cake, calf, call, came, cap, car, care, case, cash, cat, catch, caught, cave, cell, chain, chair, chat, cheap, cheat, check, cheer, cheese, chess, chick, chief, chill, choice, choose, chose, church, coach, code, coke, comb, come, cone, cook, cool, cop, cope, corps, couch, cough, cub, cuff, cup, curl, curve, cut, dab, dad, dare, date, dawn, dead, deal, dear, death, debt, deck, deed, deep, deer, dime, dirt, doc, dodge, dog, dole, doll, doom, door, dot, doubt, duck, dug, dumb, face, fad, fade, fail, fair, faith, fake, fall, fame, fan, far, fat, faze, fear, fed, feed, feet, fell, fight, file, fill, fine, firm, fish, fit, fog, folk, food, fool, foot, fore, fought, full, fun, fuss, gain, game, gap, gas, gate, gave, gear, geese, gig, girl, give, goal, gone, good, goose, gum, gun, gut, gym, hail, hair, hall, ham, hang, hash, hat, hate, head, hear, heard, heat, height, hick, hid, hide, hill, hip, hit, hole, home, hood, hook, hop, hope, hot, house, hug, hum, hung, hurt, jab, jail, jam, jazz, jerk, jet, job, jog, join, joke, judge, june, keep, kick, kid, kill, king, kiss, knife, knit, knob, knock, known, lack, lag, laid, lake, lame, lane, lash, latch, late, laugh, lawn, league, leak, lean, learn, lease, leash, leave, led, leg, let, lid, life, light, line, load, loan, lock, lodge, lone, long, look, loose, lose, loss, loud, love, loyal, luck, mad, made, maid, mail, main, make, male, mall, map, mass, mat, match, math, meal, meat, meet, men, mess, met, mid, mike, mile, mill, miss, mock, moon, mouth, move, mud, nail, name, nap, neat, neck, need, nerve, net, news, nice, niche, niece, night, noise, noon, nose, notch, note, noun, nurse, nut, pace, pack, page, paid, pain, pair, pal, pass, pat, path, pawn, peace, peak, pearl, peek, peer, pen, pet, phase, phone, pick, piece, pile, pill, pine, pipe, pit, pool, poor, pop, pope, pot, pour, puck, pull, push, put, race, rage, rail, rain, raise, ran, rash, rat, rate, rave, reach, real, rear, red, reef, reel, rice, rich, ride, ring, rise, road, roam, rob, rock, rode, role, roll, roof, room, rose, rough, rub, rude, rule, run, rush, sack, sad, safe, said, sake, sale, sang, sat, save, scene, search, seat, seen, sell, serve, set, sewn, shake, shame, shape, share, shave, shed, sheep, sheer, sheet, shell, ship, shirt, shock, shoot, shop, shot, shown, shun, shut, sick, side, sight, sign, sin, sing,

sit, site, size, soap, son, song, soon, soul, soup, south, suit, sung, tab, tag, tail, take, talk, tap, tape, taught, teach, team, tease, teeth, tell, term, theme, thick, thief, thing, thought, tiff, tight, time, tip, tongue, took, tool, top, tore, toss, touch, tough, tour, towel, town, tub, tube, tune, turn, type, use, van, vet, vice, voice, vote, wade, wage, wait, wake, walk, wall, war, wash, watch, wear, web, week, weight, wet, whack, wheat, wheel, whim, whine, whip, white, whole, wick, wide, wife, win, wine, wing, wise, wish, woke, womb, wood, word, wore, work, worse, wreck, wright, write, wrong, wrote, wrought, year, yell, young, youth, zip

A.2 Excluded words

A.2.1 Contracted forms (n=24)

he'd, he'll, he's, how'd, how's, she'd, she'll, she's, they'd, they'll, they've, we'd, we'll, we're, we've, who'll, who's, who've, why'd, why's, you'd, you'll, you're, you've

A.2.2 Morphologically complex forms (n=46)

b's, boys, buys, c's, cows, d's, days, died, dies, gays, goes, guy's, guys, guys', hoed, keys, knees, knows, law's, laws, layed, lied, lies, no's, p's, pays, rowed, rows, says, seas, sees, shall, shoes, show's, showed, shows, so's, sued, t's, tied, ties, toes, toys, twos, u's, ways

A.2.3 Function words (n=74)

been, bit, both, but, can, cause, could, did, does, done, down, five, for, four, get, got, had, half, has, have, here, hers, him, his, less, lot, might, mine, more, much, near, nine, non, none, nope, nor, not, one, same, seem, should, some, such, ten, than, that, their, them, then, there, these, third, this, those, til, till, ton, was, what, when, where, which, while, whom, whose, will, with, worth, would, yep, yes, yet, your, yup

A.2.4 Discourse fillers (n=9)

bet, feel, guess, hell, like, mean, right, sure, well

A.2.5 Interjections (n=8)

bang, damn, dude, fuck, geez, god, gosh, man

A.2.6 Words with multiple pronunciations (n=5)

lead, live, read, route, tear

A.2.7 Missing part of speech (n=46)

bing, bon, boone, byrd, cher, com, dach, dan, dat, ding, dodd, dos, fam, gore, gung, hayes, hon, hong, hyde, hype, jed, jedd, kat, kong, lan, las, los, luke, meg, neil, nerf, paul, penn, pitt, powell, reese, rhode, rom, san, schott, taj, tam, topp, wayne, whoom, yuk

A.2.8 Missing neighborhood metrics (n=31)

beep, biff, butt, cuss, foil, haul, heck, hun, jane, jeep, jock, john, joyce, ma'am, mac, med, mom, piss, psych, shawn, shit, sun, tech, ted, teen, todd, tom, whop, won, zone, zoom

A.2.9 Missing frequency (n=2)

rec, rid