A BAYESIAN APPROACH TO SPEECH PRODUCTION

by

Christo Kirov

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, MD
February, 2014

Abstract

The production of speech is one of the most widely studied topics by scientists interested in how the human mind and brain function, since it is uniquely characteristic of human cognition. Despite the everyday ubiquity of speech, the mechanisms that make it possible have proven to be extremely complex and difficult to formally model. Like other cognitive systems, the speech production system is assumed to be composed of several component stages, or levels of representation/processing. The production of a spontaneous utterance is typically described as involving first selecting a message to convey with its corresponding semantic semantic representation, encoding that message through syntax, morphology, and phonology, and ultimately creating an articulatory plan that can be used to drive the tongue and other speech organs. These different levels of processing must communicate with each other in some way, and multiple similar representations are simultaneously active during processing at each level.

Beyond this basic understanding, many of the details associated with the actual mechanisms behind these levels of processing remain up for debate and experimental results collected by linguists, psycholinguists, and neuroscientists present a number of apparent paradoxes. For example, many experiments have measured the relative effect of distractors or primes on the time required to plan and initiate a target utterance. While the tasks used in these experiments appear to differ only slightly, researchers have found that high similarity between the target utterance and the prime results in faster speech planning, and in other cases that similarity is associated with slower planning. In the absence of a strong underlying theory of speech production, such apparently contradictory results have led to complicated, somewhat ad-hoc models that focus on explaining either facilitatory or inhibitory effects, but not both. In addition, while there has been some preliminary empirical investigation of the effects of distractors and primes on phonetic variation, a systematic account explaining the range of these effects has yet to emerge.

This dissertation contributes to the empirical understanding of speech production by presenting novel experimental results describing how contextual competition in the speech envi-

ronment leads to hyperarticulation. The results of these experiments bear on which levels of processing competition takes place, and how similarity between competitors affects the level of competition. They suggest that competition occurs at specific mismatching positions between competing utterances (e.g., their onsets) rather than only at a more holistic level. Furthermore, the effects of similarity are non-linear — a competitor must differ from the target minimally in order for it to exert a measurable effect on speech production. The results also support the notion that hyperarticulation and planning latency are mechanically related. Both effects follow qualitatively similar patterns; speech seems to be hyperarticulated in just the cases when it would also take longer to plan.

In an attempt to clarify the mechanisms underlying these new results and the body of previous empirical findings, as well as to unify the formal study of speech production with that of speech perception and other cognitive functions, this dissertation applies Bayesian methods to speech production modeling. The basic hypothesis is that the levels of processing involved in speech production *communicate* with one another in the technical sense of information theory. A particular level of processing receives noisy signals from other levels indicating which representational state it should adopt. For example, the phonological encoding level receives noisy signals from a higher level that represents lexical items (lemmas). Each signal causes the receiving level to update its probabilistic belief distribution over possible representations, through the operation of Bayesian inference. If model assumptions are correct, this Bayesian decoding method is guaranteed to find the optimal interpretation of a signal given sufficient noisy samples. Formalizing the communication between levels of processing using Bayesian belief updating allows us to move towards an account of the apparent contradictions in the empirical reaction-time literature, as well as an account of the range of possible hyperarticulatory effects, with a formally simple, unified mechanism. It provides a framework that is general enough to predict the outcome of various production tasks, given knowledge of the relationships among the evidence passed between levels of processing involved, the target utterance, and the structure of competing representations.

In particular, I show how the model provides a qualitative account for pervasive patterns in priming and Stroop-like distractor studies. In many priming paradigms, target utterances are facilitated by identical primes, but slowed by non-identical, very similar primes. Depending on the relationship between targets and distractors (i.e., whether they are similar along semantic or phonological dimensions) in Stroop-like tasks, the presence of the distractors may lead to either facilitation or inhibition. I also use the model to provide an explanation for the apparent correlation between hyperarticulation and latency observed in the empirical portion of this dissertation.

Furthermore, I show that the probabilistic Bayesian approach to speech production can be extended to account for phonotactic effects that have been largely ignored in previous modeling, namely that phonotactically difficult utterances take longer to plan and are more error-prone. I also show how representing probability distributions as graphical models known as factor graphs allows active phonological processes such as syllabification and allophonic variation (and potentially higher level morphological and syntactic processes) to be included within the framework of the speech production system.

Committee members: Jonathan Flombaum, Niloofar Haeri, Akira Omaki, Paul Smolensky, Colin Wilson

Acknowledgements

I am grateful and honored to have had Colin Wilson as my dissertation advisor. Working with Colin has taught me a great deal of what I know about how to succeed as a scientist, from choosing interesting research questions to presenting my work in an engaging way. His advice and persistence, not to mention his amazing knowledge and intuition regarding both linguistic and technical matters, has made this dissertation possible.

I am also indebted to the members of my dissertation committee, and to the Cognitive Science community at Johns Hopkins University. I am particularly lucky to have had Bob Frank as the advisor for my first major project as a graduate student. It was a pleasure to navigate the perils of serial recurrent neural networks together. I thank Paul Smolensky for his kindness and continued interest in my work. It's amazing how much Paul cares about the success and happiness of graduate students. In addition, his commitment to formal and theoretical clarity has set the standard I now strive for.

I am thankful for the friendship, support, inspiration, and occasional baked goods I've received from my fellow graduate students. Life in Baltimore would have been much more difficult without them. Special thanks to the Dissertation Support Group — Erin, Dave, Mike, and Charley — for helping to keep me on track.

I would like to thank Matt Goldrick for insightful discussions about earlier versions of the work that made it into this dissertation, and for providing the foundation upon which the empirical portion of this dissertation is based.

A special mention for Adamantios Gafos, who first introduced me to the world of phonetics and phonology as an undergraduate at New York University. My career would likely have taken a very different turn without his inspiring example, and I still look forward to every chance we get to talk.

Finally, a heartfelt thank you to my parents. Their unwavering interest and confidence in my education and career, no matter where they may lead me, helps convince me that it's all worthwhile.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Overview of Speech Production

The production of speech is one of the most widely-studied cognitive mechanisms, owing to its inherent complexity and ubiquity in everyday life. Like other cognitive abilities, speech production is assumed to engage several component stages, or levels, of mental representation and processing. After several decades of systematic research, linguists, psychologists, and neuroscientists have come to a rough, if not entirely unanimous, consensus about the nature of the levels involved in speech (Levelt 1993; Griffin & Ferreira 2006; Treiman et al. 2003).

Minimally, the production of an unprompted utterance involves the following steps.[1] First, the speech production system must decide upon a semantic representation of the message the speaker wishes to convey. Next, the system must select a lexical item that corresponds to the activated meaning. The appropriate phonological form must then be selected for this lexical item. This process of phonological encoding is often broken down into 'lexical' and 'post-lexical' components. Lexical phonology involves the retrieval of phonological forms from memory. Post-lexical phonology involves the construction of phonological sequences from their component parts according to input from lexical levels and the interaction of phonological rules. Finally, a phonological plan must be used to generate a phonetic/articulatory plan that can drive the motor execution of speech. Different tasks involve other processes in addition to this basic skeleton. For example, naming a picture involves extraction of visual properties, and conversion of these properties to a semantic

[1] This description describes the process of producing a single word. Actual speech production requires planning and producing much longer utterances. It involves many additional processes that are still poorly understood and are beyond the scope of this dissertation, including selecting multiple words simultaneously, selecting appropriate morphology for each word, and arranging the words into a syntactic structure.

representation. The levels of representation/processing thought to be involved in several common

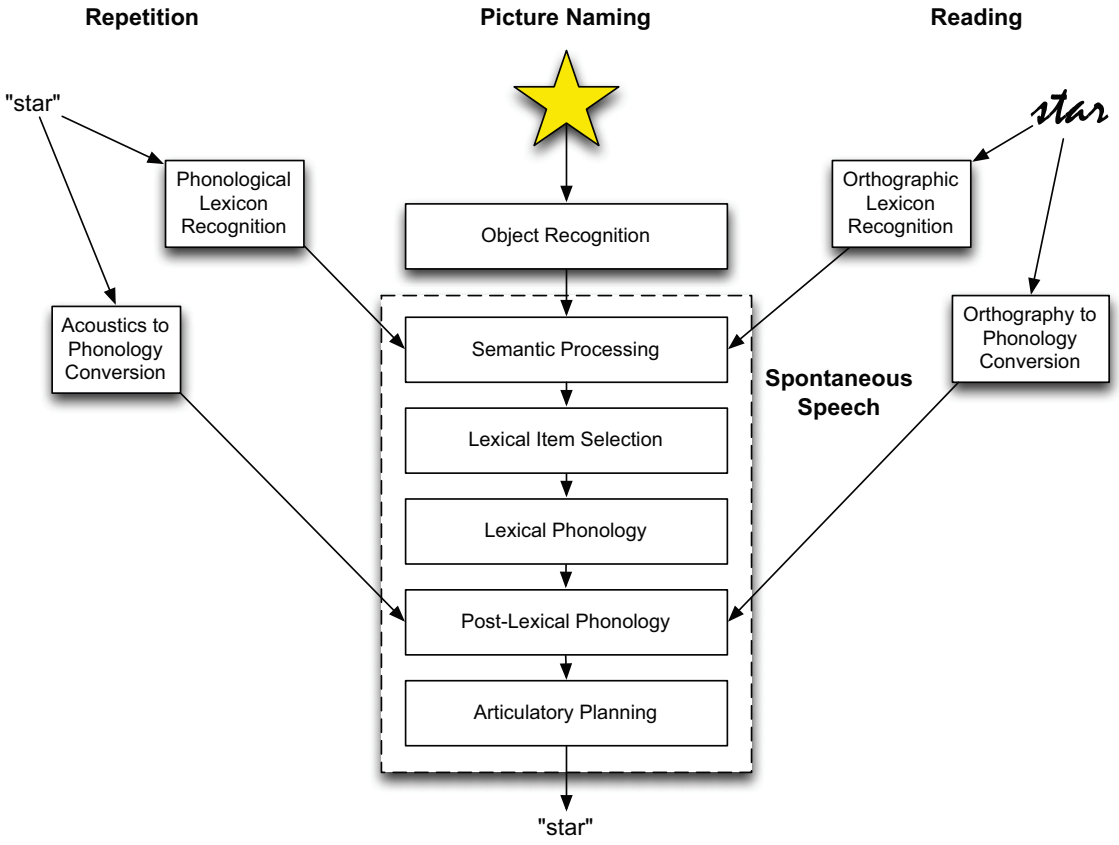

Figure 1: Schematic of speech production processes, adapted from Goldberg (2010).

While the overall architecture of the speech production system is widely agreed-upon, many of the details about the relevant representations and the interaction among levels are still debated. How should the processing that goes on within each level of processing be described (e.g., what is the format of the representations that are formed)? How do different levels of processing interact with one another? To what extent is each level engaged throughout the conception, planning, and articulation of an utterance? In abstract terms, each level is typically thought of as receiving input from one or more other levels, and forming a particular type of representation in response to that input. For example, a lexical processing level must produce a lexical item corresponding to the

semantic criteria it receives as input. However, the way in which such representations are actually formed remains a matter of research.

The goal of this dissertation is to provide insight into these questions by modeling the speech production system within the framework of *information theory* (Shannon 1948). This approach makes two related claims about the operation of the production system. The first claim is that the interaction among levels of representation/processing is formally a type of *communication*. A level receives 'messages' from other levels indicating the representational state that it should adopt. Because every kind of communication is subject to corruption by noise (e.g., noise due to the environment or internal to the operation of the mind/brain), there must be some method for overcoming conflicting and partial inputs. The second claim is that, at the functional level of description, the method used by the cognitive system is *Bayesian belief updating*. According to these claims, the operation of the speech production system is formally described as probabilistic noisy-channel communication in the technical sense defined by information theory.

Previous models of speech production, while differing in many details, have primarily been abstract neural network ('connectionist') models based on spreading activation. As a general computational framework, connectionism is very flexible, and can implement nearly any set of computational functions (Siegelman & Sontag 1995). Presumably, even the modeling approach proposed here could be implemented with an appropriately configured activation-based network. However, I will argue that restricting the behavior of interlevel communication and intralevel processing so that they conform to the rules of probability theory is appropriate for explaining many of the empirical findings associated with speech production, while minimizing model complexity and maximizing model uniformity. In other words, the information-theoretic approach adopted here allows a model equipped with a single set of computational rules to explain as large body of empirical findings. Previously, the same data had required multiple disparate and specially tuned spreading activation

models, whose behavior was often nearly as difficult to analyze and predict as the cognitive mechanisms being modeled (McCloskey 1991). Similarly, treating speech production using Bayesian modeling brings it in line with the latest developments in speech perception modeling (Norris & McQueen 2008; Norris 2006; Kinoshita & Norris 2009).

The following sections of this introductory chapter review the current body of empirical evidence that sheds light on the detailed workings of the speech production system. Models of speech production are typically compared by their ability to explain these results. Next, I present a review of the main modeling approaches that have become prevalent in speech production research, including the primary dimensions along which models tend to differ, and the shortcomings that these modeling approaches tend to have. An important contribution of this dissertation is to generalize and improve upon this earlier modeling work. Next, a basic introduction to the formal approaches used in this dissertation — information theory for describing the overall process of interlevel communication in the speech production system, and Bayesian belief updating as the mechanism that allows this communication to take place — is provided. Finally, the introductory chapter ends with an outline describing the organization of the remaining parts of the dissertation.

1.2 Review of Empirical Approaches to Speech Production

The empirical investigation of speech production has included the collection of data from both normal and aphasic populations, under both natural and laboratory conditions. The vast majority of collected data can be classified into three main types. The first type consists of the chronometric patterns associated with speech production — the varying amounts of latency required to initiate speech under different conditions. The second type of data are the patterns of production errors made by speakers under different conditions. The third type of data are patterns of phonetic variation, both articulatory and acoustic strengthening and weakening depending on the context.

One of the main factors modulating all of these effects is the interaction between competing representations within the various levels of processing involved in speech production. During the production of a target utterance, competing representations may become activated passively, simply due to their similarity to the target. During lexical/semantic processing, this includes semantic neighbors, or lexical items with a similar meaning. During phonological/phonetic processing this includes phonological neighbors, or words with a similar phonological form. Many of the studies reviewed below use 'neighborhood density,' a measure of the aggregate strength of a target utterance's neighbors in the lexicon, as a predictor variable. Clearly, some way of defining similarity between representations is necessary in order to make the notion of 'neighbor' precise.

Most of these studies focus on the phonological domain, as similarity metrics in the semantic domain are not very well developed, and it is difficult to define a cutoff point beyond which the meaning of two words is different enough that they can no longer be considered neighbors. In the phonological domain, neighborhood density is usually defined using a basic definition of the neighborhood, where a lexical neighbor is a word that differs from the target by a single phoneme substitution, deletion, or insertion.[2] When determining a word's neighborhood density using this model, all neighbors may count equally, or they may contribute according to their relative frequency compared to the target utterance (e.g., the frequency-weighted density measure in (Luce & Pisoni 1998)). However, the precise way in which the neighbor differs from the target is usually not taken into account (see Luce, 1986 for an exception). Intuitively, this may seem inadequate (e.g., "cat" seems more similar to "cad" than to "bat"), and more recent studies have examined how neighborhood structure effects speech production and recognition. Vitevitch (2002) has shown that

[2]This definition of 'neighbor' is equivalent to having a string edit distance of 1 from the target.

words with a greater proportion of neighbors sharing the initial segment are even more difficult to recognize all other things, including general neighborhood density, being equal. Davis et. al. (2005) showed that neighbors differing from a target word by deletion of a segment had a greater effect on people's ability to perform a lexical decision task on the target than other types of neighbors. Some of these effects may be due to statistics of lexical neighborhoods. De Cara et. al. (2002) noticed that neighbor types are not uniformly distributed in English; most neighbors of a given word differ in the rime[3].

The empirical thrust of this dissertation, however, is not on competitors activated because they live in a target utterance's neighborhood, but on competitors activated during the speech production process because they are an especially salient part of the current speech environment. These competitors include the set of recent utterances (words and sentences) in flowing natural speech, and any primes or distractors included by researchers in laboratory experiments. The way in which these competitors affect variation in speech latency and phonetic realization have historically been the most overlooked targets of production modeling research. In Chapter 2, novel experiments are presented which bear on which levels of processing contextual competition takes place, and how similarity between competitors affects the level of competition. As is the case for lexically-induced variation, determining which representations count as 'neighbors' is especially important. The remaining chapters of the dissertation develop a model to account for the effects of contextually-salient competitors.

[3]The rime consists of the nucleus and coda of a syllable — all the content after the initial onset. For example /at/ is the rime of /kat/

1.2.1 Chronometric Studies

Chronometric data concerns patterns in the latency associated with initiating speech production under different conditions. The time taken to initiate speech from a resting state is treated as a proxy for the time that was required for the speech production system to plan the utterance (e.g., Mooshammer 2009).

Lexical Effects

The latency associated with word production is affected by the lexical properties of the target, such as its phonotactic probability, frequency of use and neighborhood density. Phonotactic probability is usually defined according to the frequency of the one and two-segment sub-sequences that make up a word. Words composed of more common sub-sequences are deemed to have higher phonotactic probability. All other things being equal, these words are produced faster than words with low phonotactic probability (Vitevitch et al. 2004; Vitevitch & Luce 2005). Similarly, words with higher overall frequency of use are faster to produce than words with low frequency (Vitevitch 2002b; Meyer & Van Der Meulen 2000). In addition, it seems that participants are faster to name pictures of words with many lexical phonological neighbors (words from dense neighborhoods) than those of words with few lexical neighbors (words from sparse neighborhoods) (Vitevitch 2002b). Some of the attested lexical effects have proven to be controversial. In particular, it seems counter-intuitive that having many lexical neighbors should *facilitate* word production. Neighbors have an inhibitory effect in speech perception, slowing down perceptual tasks such as deciding whether a presented word exists in the lexicon or not (Luce & Pisoni 1998; Vitevitch & Luce 1999). Since other lexical properties, such as frequency, have symmetric effects on speech perception and speech production — being either facilitatory or inhibitory — the claimed asymmetry for neighborhood density is surprising. The standard explanation for this counter-intuitive finding is based on the idea that a

word is 'reinforced' by its neighbors during production via their shared phonological content (Dell & Gordon 2003; Vitevitch & Luce 1999). It is unclear why neighbors should reinforce rather than interfere with the selection of a target word. Indeed, the finding that having many neighbors facilitates speech production is not replicable in all cases, suggesting that it may depend strongly on the particular set of stimuli used in an experiment and on the task participants must perform.

For example, picture naming experiments in Spanish show the opposite effect. Words with many neighbors are named more slowly (Vitevitch & Stamer 2006; Sadat et al. 2012). English words with many neighbors that share the same onset are also named more slowly in picture naming tasks (Vitevitch et al. 2004).

Some researchers have suggested that the relevant difference between reinforcement and facilitation across studies is the strength with which neighbors become activated. Strongly activated neighbors may have an inhibitory effect, while weakly activated neighbors may serve to reinforce the activation of a target word (Chen & Mirman 2012; Mirman et al. 2010; Mirman 2011). The facilitative effect of neighbors might also be explained by phonotactic factors alone. Phonotactic acceptability is usually well-correlated with neighborhood density (Vitevitch 2002b). While the picture naming studies cited above did attempt to control the phonotactic acceptability of stimuli independently of their neighborhood density, they used measures of phonotactic probability that relied only sub-sequences of up to length two (bigrams). In actuality, phonotactic constraints can span more than two adjacent segments, and also depend on higher level structure like syllable position (Hayes & Wilson 2008). Another factor that may predict naming times but was not fully controlled was the name uncertainty associated with a particular image (i.e., participants take longer to name an image when they must decide between a number of viable names). Data from the International Picture Naming Project (http://crl.ucsd.edu/experiments/ipnp/) indicates that naming uncertainty accounts for a great deal of variability in naming times. Training participants to associate each image with a

particular name prior to an experiment, a common practice in the sort of task cited above, may not completely eliminate this uncertainty.

Contextual Effects

The time it takes to initiate speech is affected not only by lexical factors, but by contextually salient competitors in the speech environment. Generally, these effects have been studied by adding *primes* (competitors presented briefly before the target utterance or image to be named) or *distractors* (competitors presented concurrently with the target utterance or image to be named) to basic word and picture naming tasks. In some studies, competitors seem to have a facilitatory effect, speeding the production of target utterances (Gordon & Meyer 1984). In other studies, competitors seem to have an inhibitory effect, slowing production down (Meyer & Gordon 1985).

On the contextual front, a cohesive story for why and when salient competitors sometimes cause facilitation of a target utterance while causing inhibition in other tasks has yet to emerge. In Chapter 3 of this dissertations, I show how a Bayesian approach to competition in speech production is able to account for both facilitatory and inhibitory speech production effects.

Priming Producing or planning to produce an utterance can have lingering effects upon subsequent productions. Generally, these effects are stronger when the earlier production, or 'prime' is temporally closer to the second production, or 'target' (Spencer & Wiley 2008). Depending on the the exact parameters of the priming paradigm employed, and the relationships between primes and targets, target production latency may be faster or slower. However, there are several prevalent patterns. First, if the prime is *identical* to the target along some dimension of similarity under study (e.g., semantic, phonological, morphological, etc.), target production latency is minimized. If the prime is not identical to the target, then similarity often plays an inhibitory role. In various priming

paradigms, particularly those described below, the more similar the prime is to the target (without being identical), the longer it seems to take to initiate target productions.

In plan-switching paradigms, participants plan to say a prime utterance but switch to a different target when cued. This switch happens quickly when the prime and target are identical or unrelated, and slowly when they are substantially similar (Meyer & Gordon 1985; Rogers & Storkel 1998).

In masked priming, participants are shown a prime so quickly that they are not conscious of having seen it. Ferrand et. al (1996) showed then when subjects were primed with an image of word with a consonant-vowel-consonant (CVC) or consonant-vowel (CV) syllable structure, they were faster to name a target image that shared an identical syllable structure with the prime than when the prime and target shared a similar but not identical syllable structure. Unfortunately, this study did not include an unrelated prime condition.

In picture-based blocked priming studies, subjects are asked to name a series of target pictures. Each block of pictures is either homogenous (all pictures are semantically or phonologically related) or heterogenous (all pictures are semantically or phonologically unrelated). Participants are slower to name successive target pictures from homogenous blocks, whether homogeneity is defined via semantic similarity (Maess et al. 2002; Alario & Moscoso Del Prado Martín 2010; Santesteban et al. 2006; Howard et al. 2006; Damian & Levelt 2001) or phonological similarity (Mulatti et al. 2012). However, repeatedly naming the same picture (an identity condition) is fastest (Howard et al. 2006). These robust effects exist for both normal and aphasic populations (Schnur et al. 2006).

Similar blocked priming effects occur when blocks are composed of words that participants must read as rapidly as possible. In these cases, it is once again the case that repeating phonologically identical items is fast, but naming successive items that are merely similar is slower (Sevald & Dell 1994; Hilliard et al. 2011; Jaeger et al. 2012a,b; Munson & Babel 2005; Kolne 2011). An interesting aspect of many of these and similar studies is that naming targets that share an onset with

a previous production (e.g., CAT after CAR) is slower than naming targets that share a rhyme with a previous production (e.g., BAT after CAT) (Sevald & Dell 1994; Munson & Babel 2005; Wheeldon 2003)[4]. As discussed in Chapter 3 of this dissertation, this can be accounted for by considering the privileged role onsets seems to play in both perception and processing.

Stroop Tasks When a competitor is presented simultaneously with a target utterance, it is not longer referred to as a prime but as a distractor. Tasks of this form are variants of the original 1935 Stroop task, in which participants attempted to read the colors of printed words out loud, with the fact that the words themselves spelled out potentially different color names as a distraction (Stroop 1935). In the basic Stroop task, participants are slowed when target colors were presented with incongruent words (e.g., "red," presented in a green typeface). It seems that speakers cannot focus their attention sufficiently well so as to focus only on the color of a word and ignore the word itself, even though it is irrelevant to the task, and only leads to slower responses.

A common variant of the Stroop task that has been used in the study of speech production is the picture/word (sometimes picture/picture) interference paradigm. In this paradigm, participants are shown a target image whose name they must produce as quickly as they can. At different times relative to the onset of this target image (different Stimulus Onsets Asynchronies, or SOAs), they are presented with a written or spoken distractor utterance (or a distractor image embedded on the same screen as the target). Participants are told to ignore these distractors. As in the case of priming

[4]Results from a paradigm known as implicit block priming seem to show exactly the opposite result. In implicit priming, participants are taught to associate unrelated cue words with target words. During an experiment trial, they are shown a block of cue words and must name the target word associated with each one. In homogenous blocks, all the targets share their onsets or their rhymes. In heterogeneous blocks, all the targets are unrelated. Results seem to indicate that targets in homogenous blocks where onsets are shared are named most quickly, while targets in homogenous blocks where rhymes are shared and heterogeneous blocks are named equally slowly (Meyer 1991, 1990; Ardi 2004). It remains a mystery why using implicit cueing rather than displaying the target words directly leads shared onsets to have a facilitatory effect, which I will leave as an unsolved problem.

studies, several generalizations have emerged across multiple implementations of the picture/word interference paradigm.

First, participants are fastest to respond when there is no distractor (Meyer & Van Der Meulen 2000). In cases where a distractor is present, those with lower frequency tend to delay responses more than those with high frequency (de Zubicaray et al. 2012). In general, it seems that subjects are slower to name targets the more *semantically* similar they are to the distractors (Rahman & Aristei 2010; de Zubicaray et al. 2002; Mädebach et al. 2011; Mark & Wheeldon 2004; Vitkovitch & Cooper 2012). In contrast to this, increased *phonological* similarity between targets and distractors seems to facilitate production (de Zubicaray et al. 2002; Costa & Sebastian-Gallés 1998; Meyer & Belke 2007; Morsella & Miozzo 2002; Mark & Wheeldon 2004). Note that this generalization about distractors is different from the general effect of phonologically similar primes described in the previous section. This latter effect of phonological similarity has also been found in word/word interference paradigms where targets are presented as written words instead of images, or subjects are shown non-linguistic cues (e.g., hashmarks) that correspond to a particular target (Galantucci et al. 2009; Roon 2012; Roon & Gafos 2013). Again, facilitation is relative — the no-distractor case is always fastest. Any differences between different distractor types in Stroop tasks are neutralized at later SOA values, presumably because speakers have already fully prepared the production of the target by the time they see the distractor (Kleinman 2013).

Despite an overall inhibitory tendency, the effect of semantic similarity in the picture/word interference paradigm has proven to be highly variable depending on the dimensions along which similarity is measured (Spalek & Damian 2013). Thematic similarity (e.g., *mouse* and *cheese*) yields facilitation, while categorical similarity (e.g., *cat* and *dog*) yields inhibition (de Zubicaray et al. 2013). Similarly, while phonological similarity has an overall tendency to facilitate production, it can lead to inhibition in certain task variations. In particular, when a target image represents a multi-

word utterance, and distractor words are phonologically similar to words in non-initial positions of the target, both facilitatory and inhibitory effects have been reported (Jescheniak et al. 2003). Some, but not all of this variation can be explained by the idea of response criteria. If a distractor is a valid response during an experiment, its presence interferes more with target production (Mahon et al. 2007; Ardi & Piai 2013).

1.2.2 Error Patterns

Collections of speech errors have been assembled from corpora of recorded speech, from laboratory tasks designed to illicit errors from normal speakers, such as the 'spoonerism of laboratory induced predisposition' (SLIP) procedure (Vitevitch 2002b), and from cognitive testing of aphasic patients with various speech deficits (Goldrick et al. 2010). Recorded errors are typically classified as semantic, formal, or mixed (Goldrick & Rapp 2002). A semantic error occurs when, instead of an intended target utterance, a speaker produces a semantically related competitor (e.g., "lion" instead of "tiger"). A formal error occurs when, instead of an intended target utterance, a speaker produces a phonologically related competitor (e.g., "cat" instead of "cot"). Possible form errors include substitutions (replacing one segment with another), deletions (dropping a segment), insertions (adding an extra segment), perseverations (producing a segment after its intended position), anticipations (producing a segment before its intended position), and transpositions (swapping the positions of two segments). Finally, a mixed error occurs when the erroneous output is both semantically and phonologically related to the target (e.g., "rat" instead of "cat," since they only differ by one phoneme and are both animals). As discussed Section 1.3 of this introductory chapter, the relative rates of different types of errors have been used as evidence for the claim that the speech production system is 'interactive' rather than strictly feed-forward.

In general, it seems that chronometric data and error data are strongly correlated. The longer

the time required to initiate an utterance, the more likely an error is to be present in the production. For example, the lexical properties of words have been shown to affect error rates in essentially the same way as they affect speech latency. Words with high phonotactic probability are less susceptible to tongue-twisters and tip-of-the-tongue states (Vitevitch et al. 2004). Words with high frequency are similarly more error-free (Vitevitch 2002b). As in chronometric studies, high neighborhood density also appears to have a *facilitatory* effect. Words with dense neighborhoods seem to be more resistant to laboratory tongue-twister and tip-of-the-tongue error elicitation tasks (Vitevitch 2002b). Again, this does not seem to be true in cases where there are many neighbors that share the same onset (Vitevitch et al. 2004).

1.2.3 Phonetic Variation

Otherwise identical phonemes and sequences of phonemes vary in pronunciation depending on the properties of the utterance that contains them, and the speech context. This variation is often conceptualized as sitting on a scale between reduction or hypoarticulation (production of a segment or group of segments with a shorter duration and less 'extreme' articulation) and enhancement or hyperarticulation (production of a segment or group of segments with a longer duration and more 'extreme' articulations). How this scale is expressed phonetically varies from segment to segment, although duration and overall loudness are commonly used metrics for all segments. For example, voiceless stop consonant (e.g., /p/, /t/, /k/) hyperarticulation is often associated with more post-closure burst energy and longer voice onset time, or VOT.[5] Vowel hyperarticulation is often mea-

[5]Voice onset time is the time between the release of the closure during the production of a stop consonant and the initiation of the voicing associated with a subsequent vowel. See Section 2.1.2 in Chapter 2 for more information on this measure.

sured in terms of vowel space dispersion.[6] There are many established factors that affect phonetic variation in complex ways, including speech rate (Beckman et al. 2011), phonological context (i.e., the features of neighboring segments) (Gahl 2012a; de Jong & Zawaydeh 2002; de Jong 2004), prosodic context (i.e., whether the segment is stressed or not) (Cho et al. 2011), and morphological context (i.e., the class of the affix containing the segment) (Kirby & Yu 2009). The focus here, however, is on the apparent affects of competition.

In laboratory list reading tasks, words with low lexical frequency and high neighborhood density tend to be hyperarticulated. These words are produced with higher initial VOT (Goldinger & Summers 1989), higher levels of vowel-to-nasal coarticulation, or nasalization (Scarborough 2004), and expanded vowel spaces (Munson & Solomon 2004; Wright 2003). Interestingly, phonetic effects conditioned by lexical factors appear to depend only on the (frequency-weighted) density of a word's neighborhood, not on the precise phonological relationships between the word and its neighbors. For example, Scarborough (2004) found that words that were particularly confusable by their nasal consonant (i.e., had one or more lexical neighbors that differed in the position of the nasal) did not show greater vowel nasalization than words that were not similarly confusable. Words like *stem*, with minimal pair neighbor *step*, showed similar levels of nasalization as words like *plank*, with no nasal-differing neighbors in the lexicon. Similarly, Goldinger & Summers (1989) found more VOT enhancement for voiceless-initial words from dense neighborhoods than those from sparse neighborhoods, even though both sets of words had exactly one minimal pair lexical neighbor that began with a voiced sound.

[6] Vowel space dispersion, also referred to as expansion, is calculated as the mean euclidian distance of a set of vowel tokens from the center of the speaker's vowel space, the mean F1/F2 values over all the speaker's vowel tokens (Bradlow et al. 1996).

Contextually induced competition also affects phonetic realization, apparently inducing hyper-articulation when present. For example, Buz & Jaeger (2012) found that word duration in a corpus of running speech is negatively correlated with distance to the nearest previously mentioned neighbor: neighbors mentioned in the recent past, against which the current word must plausibly compete, condition longer phonetic realizations. However, when a word is repeated multiple times in running speech, each repetition tends be more reduced, suggesting that words do not compete with themselves (Fowler 1988; Lehnert-LeHouillier 2010). Baese-Berke & Goldrick (2009) found VOT lengthening for voiceless-initial target words in the context of voiced-initial neighbors (e.g., the word CAP in the context of the word GAP). Tilsen (2007,2009,2013) found that target vowels productions tended to have more extreme formant values after speakers planned to or produced a related prime vowel.

Phonetic Variation and Speech as Part of an Optimal Communication System

A popular perspective on speech, and language in general, has been that many aspects of its structure and performance are functional in nature. That is, if language is to be a successful means of communication between speakers, it might be organized in such a way to make communication particularly efficient. Intuitively, language should be organized so that complex messages can be conveyed in as little time as possible without many errors or misunderstandings. This intuition is conveniently formalized in the language of information theory, which was developed to describe communication systems mathematically (Shannon 1948). In information theory, all communication occurs through noisy channels, which distort the data that pass through them. A communication system has a maximum rate at which it can transfer information with negligible errors, known as the channel capacity. In order to approach this optimal rate individual messages should be as short as possible (compression), but still retain enough information (redundancy) so that the person on the

receiving side of the system is able to recover the original intended message despite knowing that the raw data they received has been partially corrupted by the noisy channel.

There is evidence that as languages evolve over time, they become more efficient according to information-theoretic criteria. One place where this diachronic pressure can be seen is in the structure of lexicons. information theory places a limit on the number of possible words of a certain length. For example, as more and more single phoneme words are added to the lexicon, they necessarily become more similar to each other since there are only so many distinct sounds a human can produce. To avoid confusion, it eventually becomes necessary to add longer words to the language (Plotkin & Nowak 2000). To minimize the length of messages, it pays to structure the lexicon so that these longer words are less likely to be used at any given time. The negative log of the probability of a word given the preceding context is referred to as it's information content. It has been been shown that the length of words in the lexicon of English correlates well with their average information content (Piantadosi et al. 2011, 2009).[7] A concept related to information content, functional load, has been used to explain diachronic changes in the sound patterns of languages (Surendran & Niyogi 2006; Wedel et al. 2013). For example, if the words of a language are primarily distinguished by the quality of their stressed vowels, then vowels carry a high functional load. In such a language, vowel sounds are unlikely to merge in stressed positions, but might merge in unstressed positions where distinguishing between different vowels doesn't help distinguish between different words.

Evidence that languages become more efficient over time also comes from the literature on

[7]This is consistent with a finding in information theory stating that source signals are most compressed when each symbol to be transmitted has a representation size proportional to the negative log of its probability (Cover & Thomas 2006). However, we know that natural language cannot be maximally compressed. It must remain sufficiently redundant to prevent errors.

language learning. There is evidence that as new learners acquire a language, they spontaneously reorganize it at a grammatical level so as to make it more efficient. In an artificial language learning experiment, subjects learned to apply optional case markings to nouns only when their grammatical role was ambiguous, rather than randomly as in the training data they were presented with. The learned language was thus clear but free of extraneous case markers (Fedzechkina et al. 2012).

Another set of corpus-based approaches has examined the phonetic variation associated with speech in relation to its information content. Unlike the studies of overall lexicon structure described above, which look at the length of words in response to their *average* information content, these studies look at how particular utterances at different size scales from the phoneme level up vary instance-by-instance according to how probable (i.e., predictable) they are. Segments (van Son & Pols 2003, 2002; van Son et al. 2004; Everett et al. 2011; Clopper & Pierrehumbert 2008; Cohen-Priva & Jurafsky 2008), syllables (Aylett & Turk 2004), and words (Tily et al. 2009; Demberg et al. 2012; Dilts et al. 2011; Lam 2012; Kuperman & Bresnan 2012; Wiener et al. 2012; Shaw 2012) that are less predictable in context[8] are longer in duration or less reduced than more predictable segments and syllables. The result showing that repeated words tend to be more reduced (Fowler 1988; Lehnert-LeHouillier 2010) can also be treated as an instance of this phenomenon, insofar as producing a word once usually makes it more likely to be produced later on. Some researchers have noted that since items that have more information content usually have longer duration, information content tends to be spread out more evenly per unit of time. Thus, this set of phenomenon has been referred to as uniform information density (UID) or smooth signal redundancy (Jaeger 2010; Frank & Jaeger 2008; Levy & Jaeger 2006; Qian & Jaeger 2010; Genzel & Charniak 2002; Keller 2004;

[8]Often determined by simple probabilistic models applied to corpora of running speech.

Manin 2006). UID implies that the efficiency pressure that ultimately resulted in the structure of lexicons applies on a much shorter timescale.

These studies suggest that pressure to achieve efficiency exists in moment-to-moment speech performance, resulting in phonetic variation. Lindblom characterized speech production as a balance between the speaker's desire to conserve effort (i.e., to affect compression), and their desire to accurately convey their message (i.e., to retain sufficient signal redundancy) (Lindblom 1990). According to his H&H (Hyper- and Hypoarticulation) framework, speakers hyperarticulate when they need to ensure accurate reception of their message in potentially adverse speaking conditions and hypoarticulate when they can as long as listener understanding can be maintained.

A number of high-level systemic changes to speech production seem to fit within this framework. Child directed speech, and speech directed at the hard-of-hearing, fall within a mode of production known as clear speech. When operating in this mode, speakers tend to speak more slowly, more loudly, and tend to hyperarticulate more (Uchanski 2008; Amano-Kusumoto & Hosom 2011). For example, clear speech tends to show increased vowel dispersion (Bradlow et al. 1996). Natural clear speech recorded in actual communicative settings seems to be more intelligible than ordinary speech when played back to listeners (Bradlow & Alexander 2007; Smiljanić & Bradlow 2009, 2005), suggesting that speakers are employing a successful set of strategies to increase recognition accuracy.[9] A similarly listener-oriented mode of speech is Lombard speech, the set of modifications speakers make in response to loud ambient noise in the speech environment (e.g., at a loud concert). The Lombard reflex characterized by increased overall volume and

[9]However, Scarborough & Zellou (2012) show that 'pretend' clear speech, recorded from participants who were told to pretend they were speaking to someone hard-of-hearing, was less intelligible than actual clear speech, suggesting that speakers' conscious knowledge of what constitutes clear speech may not be optimal.

pitch (Lau 2008; Cooke & Lu 2009; Welby 2006; Summers et al. 1988; Zhao & Jurafsky 2009). Simply by virtue of increased volume, moderate Lombard speech is easier to understand in noisy environments. However, if the noise level increases beyond a threshold, speakers tend to begin shouting, which once again decreases intelligibility.

The modifications made by speakers during the production of clear or Lombard speech are assumed to be systemic, affecting global speech parameters such as overall pitch or volume (Hazan & Baker 2011). However, the more specific phonetic variation effects discussed above have also been amenable to a functional analysis. Empirical evidence indicates that words with low frequency and high neighborhood density are more difficult for listeners to recognize, presumably due to competition (Luce & Pisoni 1998). This has led to the characterization of these words as 'Hard.' 'Easy' words, on the other hand, have high frequency and few neighbors (Wright 2003). The H&H framework suggests that speakers should hyperarticulate precisely those utterances that would be hard to recognize. Thus, it makes sense from a functional standpoint to increase the initial VOT of Hard words, increase their levels of coarticulation[10], and increase their level of vowel dispersion. However, it should be noted the status of these lexical effects, particularly as instances of listener-oriented optimization, has recently been challenged, particularly with respect to vowel-space dispersion. It seems that when phonological context is precisely controlled, the difficulty associated with recognizing a word no longer predicts its level of vowel dispersion (Gahl 2012a). Furthermore, lexical frequency and neighborhood density seem to have inconsistent effects in running speech, as opposed to in single word productions. For example, words with high neighborhood density tend

[10]It may seem intuitively that greater degrees of coarticulation 'smear' the acoustic signal, making words less clear. However, more coarticulated speech has been shown to be more intelligible than less coarticulated speech (Scarborough et al. 2011). Presumably, coarticulation provides redundant information about the segments that make up an utterance, giving listeners extra information about which segments are yet to come as they process an earlier part of the signal.

to be hypo- rather than hyperarticulated in running speech (Yao 2010; Gahl 2012b), at odds with the results presented above. One possible explanation for this is that during running speech, the predictability of various words is always in flux as the context changes, possibly changing the set of active competitors in a target word's neighborhood. Raw measures of lexical frequency and density would then be valid when in the special case of single word production, when no immediate context is available and speakers rely on their aggregate knowledge, but might not necessarily be valid during running speech. Indeed, Beattie (1979) found that dysfluencies in word production were always predicted by contextual probability, but were only predicted by raw lexical frequency to the extent that this was correlated with contextual probability. Similarly, Scarborough et al. (1977) found that independent frequency effects in word reduction disappeared when context was taken into account.

Contextual phonetic variation can be motivated in a way similar to that of phonetic variation conditioned on lexical factors. If there is a particularly salient competitor in the speech environment, either a recently uttered neighbor in natural speech or a deliberate distractor as in Baese-Berke & Goldrick's experimental study. The VOT study is particularly helpful for understanding how hyperarticulation might serve to enhance the linguistic signal in such a way that errors become less likely. In general, hyperarticulation should have the effect of making a target utterance more different from its competitors. If a target utterance (e.g., "pig") is presented in the context of voiced-onset competitor ("big"), hyperarticulating "pig" by lengthening its initial VOT serves to make the initial /p/ phoneme more different from the the initial phoneme of "big," which has very short VOT. The Baese-Berke & Goldrick study serves as the foundation for the empirical investigation presented in Chapter 2 of this dissertation, and is presented in more detail in Section 2.1.1.

While it may be fruitful to analyze the potential efficiency advantages conferred to speech by various phonetic variation effects (Kirov & Wilson 2012; Flemming 2010), this purely functional approach does not describe the mechanisms behind the effects, only the high-level motivation be-

hind them. In the case of moment-to-moment synchronic effects (as opposed to long-term pressures on lexical structure and grammar), some researchers have proposed that speech production includes mechanisms that predict how listeners would respond to a particular utterance, and optimize the production of the utterance accordingly (Jaeger 2013; Jaeger & Ferreira 2013). There is evidence that speakers can predict what their utterances will sound like, and that the speech production system constructs speech using acoustic/perceptual rather than articulatory goals. When their articulators are perturbed by an external device (e.g., their jaw is pulled down or their tongue is depressed), speakers can alter their articulations on-the-fly so as to minimize changes to the acoustic signal (Guenther et al. 1997). In addition, it seems clear that high-level systemic effects such as the shift to a clear speech mode of production require the speaker to at least notice that they are talking to a child or someone hard-of-hearing. Furthermore, speakers seem to have accurate subjective ratings of how intelligible words will be under noisy conditions (Bowdle & Wright 1998). Some speakers also have a tendency to match the speech patterns of their interlocutor over time, potentially leading to greater mutual intelligibility (Hazan 2012).[11] However, it is unclear to what extent such listener-modeling mechanisms are needed to explain the various effects discussed in this introductory chapter. In particular, many of the words that are difficult to perceive in a given context, may also be difficult to produce due to competition effects. To the extent that speaker and listener difficulty are correlated, the speaker does not need an extra mechanism to guess the listener's state. Even without such an explicit mechanism, the speech production system may still produce speech that is well-suited for recognition by the listener (Bradlow 2002; MacDonald 2013). It should also be noted that many of the effects described above were recorded in laboratory experiments in which

[11] For their part, listeners also adapt to novel voices over time, improving their own understanding (Dahan et al. 2008; Kraljic & Samuel 2007).

no listener was involved, meaning that the speech production system would have had no subject to model.

In this dissertation, I focus on modeling the possible mechanisms behind low-level phonetic variation effects. The model presented does not include high-level listener state prediction components or self-monitoring components, in an effort to explain the maximum amount possible while keeping the model simple. This does not put it at odds with the functional hypothesis that speech is one part of an overall efficient communication system, since the functional approach is agnostic with respect to the actual mechanisms that produce efficient speech. In future work, the extent to which an explicit listener-modeling component is required could be explored experimentally by directly manipulating speakers' beliefs about listeners' knowledge state on a trial by trial basis.

1.3 Review of Modeling Approaches to Speech Production

To begin, the modeling of speech production has thus far focused on the planning and production of single words (although previously produced words may be allowed to affect these processes). The new modeling approach presented in this dissertation is similarly limited in scope. Understanding the full range of long-distance interactions (e.g., the possibility that some sequences of words may be planned and produced as a single chunk) is left for future research.

While having in common with the model proposed here the idea that speech production is broken down into multiple levels of processing, most previous models of single-word production have been based on the idea of spreading activation in an artificial neural network, and thus fall into the category of 'connectionist' models. Levels of processing in connectionist networks contain collec-

tions of nodes, one for each possible representation in the pool[12]. For example, a lexical level of processing in a connectionist network may contain one node for each stored lemma representation. Nodes may be connected to other nodes in the same processing level, or to nodes in other processing levels. These connections serve as pipes through which activation can spread between different representations. Numerical weights on the connections determine how a node's activation affects that of its neighbors. The weight from one unit to another can be positive (facilitatory) or negative (inhibitory).

At a high level, connectionist models of this sort are used to model the computations involved in speech production as follows (albeit with many model-specific differences, as discussed below). The target representations in a particular level of processing (.e.g., the lexicon of known phonological forms) receive input activation from connections to external levels. Activation is allowed to spread within the network until a selection criterion is met (either a set time limits passes, or some representation node reaches a pre-defined activation threshold). At this point, the level of processing can be said to have 'decided' on a particular representation for production, typically the representation whose form has the highest activation.

There are several important dimensions along which the computational skeleton described above can be customized, and which define the space of theories of speech production that connectionist models instantiate. The most important of these dimensions is the type of interactivity between different levels of processing.

Interactivity refers to the passing of activation between different levels of processing while one

[12]Identifying each representational state with a single node is referred to as a 'localist' representation. It is also possible to encode the identity of a state by a pattern of activation across multiple nodes. This alternative is referred to as a 'distributed' representation. Since the most common speech production models use localist representations, I will only focus on them here. However, for the utility of distributed representations in language modeling, see Smolensky et. al (2006).

or both are still undergoing a decision process. In a fully discrete model, each level of processing must decide upon a representation before competition can begin in a subsequent level (e.g., the selection of a lexical item is completed before phonological encoding can begin). There is some neuropshysiological evidence that speech production has this discrete sequential structure. Intra-cranial electrode signals from Broca's area seem to indicate three temporally ordered peaks of activation roughly corresponding to lexical access, grammatical encoding, and phonological encoding (Sahin et al. 2009) (but see Rahman, 2003 for EEG evidence of parallel semantic and phonological processing). Behavioral evidence, however, seems to favor a more interactive account. If a model allows a level of processing still undergoing a decision process to spread activation to another level, it is referred to as including *cascading* activation. If the module receiving the cascading activation is allowed to send activation back, the is said to contain *feedback*. Feedback allows the computations occurring in downstream processing levels to affect decisions being made by upstream levels. Evidence for the presence of cascading and feedback has primarily come from an analysis of error patterns.

If levels of processing were entirely discrete, then errors made at one level and errors made at subsequent levels should be statistically independent. The rate of observed joint mixed errors (those including both a semantic and phonological error) should be equal to the product of the individual rates of semantic and phonological errors. In reality, it seems that mixed errors are overrepresented (Goldrick & Rapp 2002). Similarly, phonological errors display a real word bias — errors are more likely to be existing words in the language. This would not be expected unless lexical levels of processing were interacting with phonological encoding (Moat 2010).

The next two sections review the two most comprehensive and widely-cited connectionist models of speech production — Dell's (1986) interactive activation model and Roelofs' WEAVER++ model.

1.3.1 Dell's (1986) Model and Derivatives

Dell's original (1986) model of word production was primarily designed to explain error patterns in word production. It includes a number of potential levels processing, including levels representing lemmas, morphemes, syllables, phonemes, features, etc. (exactly which levels are actually used at any one time depends on the phenomena being modeled). Representations in a level are connected to their associated representations in another level (e.g., a phonological word form is connected to each of its component phonemes). There are no links between nodes in a level, only across levels, and all connections are positively weighted — there is no inhibition in the network. Production planning starts by introducing a jolt of activation to an appropriate set of source nodes (e.g., a set of semantic primitives), and proceeds via fully interactive (including both cascading and feedback) activation spreading. As a result of feedback, similar representations in the same level may be simultaneously activated via their shared connections on another level (e.g., similar phonological forms are co-activated via their shared phonemes). Usually, some level of noise is added to the activation as it spreads. Planning ends after a fixed amount of time has passed, and the most highly activated nodes in each level are selected for production.

By varying the amount of noise in the system, the model is able to account for a wide range of error patterns, including the real word bias effect discussed above (Moat 2010). By tuning the relative levels of cascading and feedback in the network, it is also able to account for the distribution of mixed semantic/phonological errors (Goldrick & Rapp 2002).

However, there are a number of drawbacks stemming from the design of the model that have limited its use beyond modeling of error patterns. First, because the planning process always runs for a fixed number of time steps, the basic model cannot account for chronometric variation in speech production. While the model can in principle include levels of processing for the selection

of individual phonetic features, and the numeric level of activation associated with each feature might somehow be used to model phonetic variation along a continuum (e.g., varying duration or voice-onset time), the model has yet to be used in this way. Finally, simulations involving Dell-style models have been limited to small networks — usually involving lexicons of less than 20 words — in order to ensure interpretable model behavior. This puts into question how well such models scale to more realistic network sizes (Chen & Mirman 2012).

More recent work has addressed some of these drawbacks within the limited domain of lexical neighborhood interactions during speech production. Chen & Mirman (2012) used a variant of Dell's original model that made decisions not by selecting the most active node after a fixed amount of time, but by selecting the first node to reach an activation threshold. This allowed the model to make chronometric predictions. In addition, nonlinear inhibitory links were added between different word form representations (the inhibitory links were nonlinear in the sense that they only became open once a connected word form passed an activation threshold). This allowed strongly activated neighbors to compete with each other. Both of these modifications were necessary in order to provide an account of why neighbors sometimes seem to have a facilitatory effect on single word production, while at other times inhibiting it (see above for a review of the relevant results). However, modified Dell-style models of this sort have yet to be tested across a wider range of chronometric phenomena.

1.3.2 WEAVER++

Unlike the Dell model above, Roelofs' (1997) WEAVER (Word Form Encoding by Activation and Verification), eventually upgraded to WEAVER++ (Levelt et al. 1999), based on Levelt's (1993,1999) theory of speech production, was designed primarily to account for chronometric data — particularly the reaction times associated with various Stroop tasks. It also adds a number of additional

mechanics to the basic spreading-activation template.

Unlike in Dell's model, similar representational nodes within levels are directly connected to each other by facilitatory links[13]. Processing within a level consists of sending activation to a target node and allowing it to spread across any of the connections within the level. Normally, the target node will eventually cross an activation threshold. However, this does not determine the time required to select the node. Instead, activation continues to spread and a probability distribution over selection times is defined based on the ratio of the activation of the target node and all other nodes:

$$P(\text{access target } m \text{ at time } t) = \frac{a(m,t)}{\sum_i a(i,t)}$$

where i indexes over all competing representations. If the ratio is high (i.e., the target is considerably more active than its competitors), it will be selected quickly. This additional mechanism allows the model to achieve competitive effects and describe chronometric data without including inhibitory links. Levels of processing in WEAVER++ are fully discrete. Once a representation is selection on one level, target nom/des that correspond to it are marked for selection on downstream levels.

WEAVER++ has been successful at accounting for the overall facilitative effect of phonologically similar distractors (relative to phonologically distant distractors) in the picture/word task. Initially, the model marks the phonological form of the target for production. However, before activation spreading can complete at the phonological level, the distractor is presented and also marks its form for production. Since this happens later during the spreading activation process, the target form will likely reach its activation threshold first and be selected for production. However, the activation ratio that determines the reaction time associated with the target will differ depending

[13]In Dell's model, similarity effects arise from feedback and shared components; similarity in WEAVER++ must be explicitly stipulated via the choice of connections within a representational level.

on how similar it was to the distractor. If the distractor and the target are *not* phonologically similar, then activation from the distractor node will spread to many of the target's competitors and the denominator in the ratio will be large, resulting in a slower reaction time. If the distractor and the target *are* similar, activation will still spread to the target's competitors, but will also spread to the target itself. Thus, the numerator in the ratio will also increase, balancing out increases in the denominator and resulting in a relatively faster reaction time.

Like Dell's model, WEAVER++ has some drawbacks. Since it is fully discrete, it cannot accurately describe the full range of error rate patterns Dell's model describes (Goldrick & Rapp 2002). In fact, the model explicitly includes mechanisms that preclude errors, as a response to evidence that error rates tend to be low and vary less than reaction times (Levelt et al. 1999). First, activation spreads in the model without any noise. Second, there is an external verification mechanism that checks if selections made at a downstream level correspond correctly to earlier selections at higher processing levels. In order for errors to occur, the selection mechanism must explicitly be made to break down (usually at a stipulated rate), and verification must fail (again, at an ad hoc rate). In its current formulation, WEAVER++ also cannot handle phonetic variation. The model assumes that speech is produced by stringing along a number of pre-built syllable programs. Phonetic features are not allowed to vary and cannot be selected individually. There is evidence that this is not the case. An articulatory task where participants were forced to cut off speech at unpredictable times shows that they can drop or add individual features to their production plans as needed (Tilsen & Goldstein 2012).

1.3.3 Alternative Models for Specific Tasks

There are a number of additional models that are limited in scope to specific aspects of speech production, and do not provide detailed accounts of the interactions between different levels of process-

ing. Several models rely on a mix of connectionism and dynamical systems theory to examine just the coordination and motor execution of articulatory gestures. These include Saltzman & Munhall's (1989) Task-Dynamic model and Guenther's (1995) DIVA model. Roon (2012,2013) discusses a model of gestural selection based on dynamical field theory (Erlhagen & Schöner 2002). This model was designed specifically to address the reaction time results associated with the phonological cue-distractor tasks described above (participants are faster to name a target word associated with a non-linguistic cue if simultaneously presented with a phonologically similar distractor as compared to a phonologically distant distractor). It does not explain the inhibitory effects of phonological similarity seen in other tasks. Goldrick (2008) suggests that a harmonic grammar can account for hyperarticulation if highly activated inputs to the grammar correspond to stronger faithfulness constraints. This approach does not explain why or how inputs to the grammar would be more highly activated in cases where they are hyperarticulated. The chronometric results described above are also beyond its intended scope.

1.4 Basic Principles of information theory and Bayesian Methods

As mentioned above, the approach to modeling speech production adopted in this dissertation is grounded in information theory. It starts with the assumptions that the system is composed of several levels of processing, and that the interactions among these levels can be described as sending messages through noisy communication channels. As in the children's game of "telephone," the intended messages are usually partially altered by the time they reach their destinations. information theory describes the limits on how much information can be recovered from corrupted messages, as a function of the probabilities describing how the messages are likely to change as they pass through the noisy channel. Receiving levels must act as *decoders* that extract as much of this information as possible. A potentially optimal decoding strategy, particularly when it is known ahead of time what

messages are likely to be sent, is Bayesian belief updating (Cover & Thomas 2006).

Bayesian methods allow the principled combination of prior beliefs and new evidence to form optimal inferences about data. At their core lies Bayes' Theorem:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

This states that the probability of a hypothesis given new evidence $P(H|E)$ is equal to the probability of the evidence given the hypothesis $P(E|H)$ (a term referred to as the likelihood) times the probability of the hypothesis prior to receiving the evidence $P(H)$, divided by the overall probability of the evidence $P(E)$. The $P(E)$ term is defined as:

$$P(E) = \sum_{i=1}^{i=n} P(E|H_i)P(H_i)$$

This theorem can be applied repeatedly every time a new piece of evidence is found. After every update, the posterior probability of a hypothesis $P(H|E)$ becomes the prior $P(H)$ for subsequent updates.

Within Bayesian reasoning, probabilities are interchangeable with beliefs. Or rather, degrees of belief conform to the axioms governing probabilities. Thus, by calculating the posterior probability of a hypothesis using Bayes' theorem, we are also changing our beliefs about how likely that hypothesis is to be true. Assuming that our initial prior beliefs about the set of possible hypotheses and our likelihoods are correct, then Bayes' theorem represents a mathematically optimal way to update our beliefs (Norris & McQueen 2008; Norris 2006). It is this potentially optimal performance that makes Bayesian methods an important tool in the study and implementation of communication systems as formalized by information theory. When a level of processing receives corrupted messages, it can use Bayes' theorem to combine its faulty input (the evidence) with any prior beliefs about what the intended message probably was, and make a rational inference indicating the most likely intended message given the evidence.

Of course, Bayes' theorem only provides an optimal way of using prior beliefs. If those prior beliefs are wrong, then any inferences derived by Bayesian methods may also be wrong. This is both a valuable feature of Bayesian methods and a source of heavy criticism. It is a feature since prior information is often available in some form, and usually leads to better inferences in practice, so it is good to have an optimal way of using it. It is a source of criticism since prior beliefs usually have a subjective component, and thus might result in biased inferences.[14]

It is important to note that Bayesian statistical modeling may be used to analyze language data at multiple levels. Commonly, Bayesian modeling is used to describe the sources of variance in experimental data. In this case, it serves as an alternative to or generalization of traditional methods of data analysis, particularly generalized linear models (e.g., ANOVA). VanDam & Silbert (2010), for example, use such an analysis to search for variables that accurately predict voice onset time duration. In this usage scenario, Bayesian modeling is used to describe observed distributions of experimental data, but Bayesian mathematics are not necessarily assumed to have any part in generating the data in the first place.

In this dissertation, I follow a more recent trend in Bayesian modeling in which Bayesian inference is assumed to be part of the mechanical workings of the mind/brain (Norris & McQueen 2008; Norris 2006; Sanborn et al. 2010; Stocker & Simoncelli 2006; Simoncelli 2009). That is, the brain itself may make inferences according to Bayes' rule (or, more likely, an approximation to it). If that is the case, then experimentally observed behaviors may follow as a mathematical consequence of the restrictions on calculations enforced by sticking to the probabilistic framework Bayes' rule provides.

[14]In practice, researchers have discovered types of prior beliefs work well in various situations. This includes so called 'uninformative' priors that are designed to minimize bias (Jaynes 1968).

1.5 Structure of the Dissertation

The remainder of the dissertation is organized as follows. In Chapter 2, I present six experimental studies examining how competing words in the speech environment affect the planning and phonetic realization of words. The overall set of results suggests that salient competitors in the speech environment affect the planning and production of a target utterance as long as they are sufficiently similar to the target. Competitors (most appropriately primes, according to the prime/distractor dichotomy described above) that differ from the target by approximately one phonological feature slow down speech planning and may cause hyperarticulation of the target. These effects of word competition appear to drop off rapidly as the phonetic/phonological distance between targets and competitors increases. These results are important for setting up the theoretical foundations for the modeling presented in later chapters, and for motivating the parameterizations of the models developed.

In Chapter 3, I describe a novel approach to modeling speech production based on Bayesian statistics and information theory. The model maintains the levels of representation/processing traditionally associated with speech production, including lexical selection and phonological encoding, but asserts that 'messages' sent from one level to another are treated as noisy evidence for a particular representational state. Each level of processing maintains a probability distribution over possible representational states, and uses Bayes' rule to update this distribution in accordance to the evidence received. The goal is to provide a uniform and formally principled alternative to the activation-spreading network mechanics typically used to model speech production. I show how this approach can be used to explain a number of experimental results relating to speech latency and phonetic variation. In particular, the model can account for the generalizations described in the sections on Priming and Stroop-like tasks above, including the fact that similarity between primes/distractors

is sometimes facilitatory and sometimes inhibitory. It also accounts for the relationship between planning latency and hyperarticulation discussed in Chapter 2, something which no previous model has attempted to explain.

While the basic Bayesian model presented in Chapter 3 is sufficient to explain many empirical phenomena, it is hampered by a number of simplifying assumptions about the the nature of representations in the speech production system. These simplification preclude the model from providing explanations of phenomena such as the production of novel unknown words, and the apparent effects of phonotactic knowledge on the production of speech. Furthermore, the simplified model cannot describe how productive processes in syntax, morphology, and phonology are realized within the speech production system. In Chapter 4, I extend the description of the model to show how these processes could be treated. I provide a proof-of-concept implementation showing how phonotactic knowledge in particular can be incorporated into the model. The implementation is capable of capturing the fact that phonotactically dispreferred utterances take longer to produce.

Finally, Chapter 5 provides a summary of the dissertation, and suggests future avenues of research for the Bayesian modeling approach.

2 Empirical Investigation

2.1 The Relationship Between Contextual Competition and Phonetic Variation

The empirical results reviewed in the introductory chapter indicate that the presence of salient competitors in the speech environment affects the phonetic realization of utterances (Buz & Jaeger 2012; Baese-Berk & Goldrick 2009). In particular it seems that higher levels of competition lead to hyperarticulated productions. However, these studies have left a number of unanswered questions about the nature of this competition. How must competitors be related to the target utterance in order to influence its production? Are all phonological neighbors (by the standard definition, a form differing by any one segment from the target) sufficiently strong competitors, or should the set of strong competitors be defined more precisely? Is competition-induced hyperarticulation limited in scope to just those segments of a target utterance that differ from its competitors?

The six studies presented in this chapter provide preliminary answers to some of these questions, and in turn bear on the theoretical foundations of the speech production system that will be modeled in later chapters. Two theoretical dimensions are of particular interest. First, it is commonly assumed that there are at least two levels of phonological processing within the speech production system. A lexical level retrieves full stored phonological forms for words. A post-lexical position-based lower level must select individual segments to fill positions in a phonological string. This lower level is necessary for the production of novel word forms that are not yet stored in the lexicon, but it is likely that the lexical level also serves as input to it. In principle, competition could occur at either or both of these levels. Competition at the lexical level should involve whole-words. That is, the effects of a competitor should be visible across the entire target form, regardless of the precise location of any similarities or differences between the target form and the competitor. On the other hand, competition at the post-lexical level should be position-specific. The effects of a

competitor would only be visible in those positions in the target that differ from the competitor.

Second, even having pinpointed the locus of competition effects, a suitable similarity metric is required to determine exact strength and nature of these effects. One possibility is that the similarity metric is very broad, as in the common definition of phonological neighbors (i.e., any form within one-segment of the target is sufficiently similar to it to count as a neighbor and viable competitor). Another possibility is that the metric is very narrow (e.g., differing by an entire segment is too much — only forms that differ by just a feature or two are sufficiently similar to the target to cause strong competitive effects).

The studies presented in this chapter are limited to examining competition between phonologically, or formally, related words. Competition between semantically related words is left for future research, as semantic similarity is not well-defined in the current literature.

2.1.1 Experimental Paradigm

All hyperarticulation experiments (Experiments 1 through 5) used an experimental paradigm adapted from Baese-Berke & Goldrick (2009). The goal of the paradigm is to simulate a communicative situation in which confusion is possible, though unlikely. The paradigm involves two participants, one in the speaker role and one in the listener role. Both participants sit at their own computer screen. The listener cannot see the speaker's screen but can hear the speaker talk. In each trial of the experiment, two or more words appear on both screens — a target word along with competitor words that may be similar to the target. After 1500ms, the target word becomes highlighted on the speaker's screen, and they are obliged to say it out loud. At this point, the listener clicks the word they hear, attempting to match what the speaker said. The speaker's pronunciation of the target is recorded and can be analyzed acoustically. The overall setup is shown in Figure 2.

This paradigm has the advantage of being able to precisely control a target word's experimen-

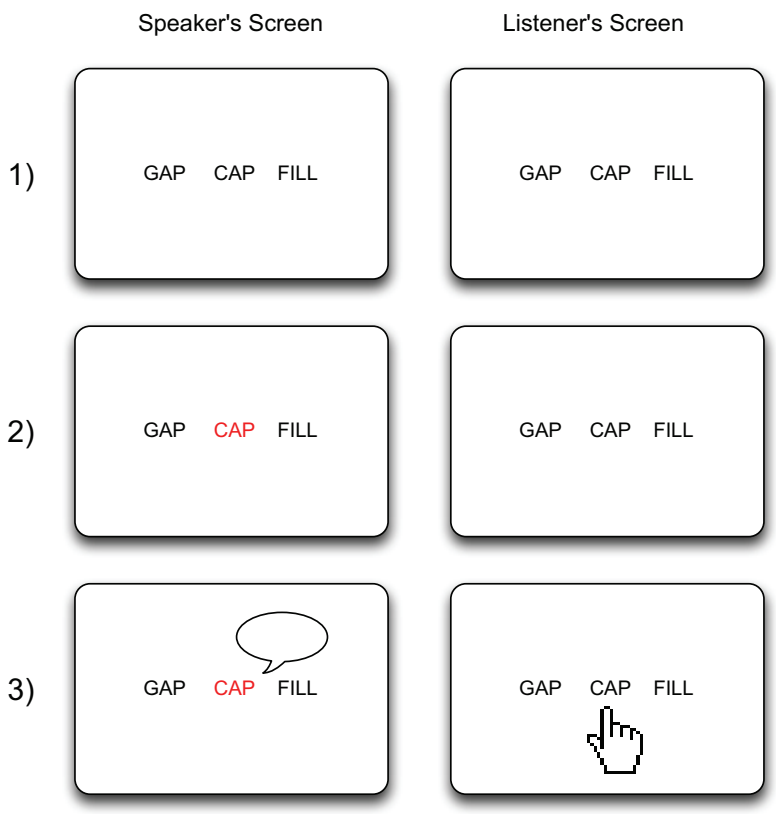

Figure 2: Experimental paradigm. Initially, both speaker and listener screens show the same three words. After 1500ms, the target word becomes highlighted on the speaker's screen. At this point, the speaker must say the target out loud, and the listener must click on the word that was heard.

tal 'context' (the competing words that appear on-screen with it) and includes motivation for the speakers to communicate clearly, as they are made aware if the listener does not click on the target word. However, the paradigm involves an artificial laboratory task and is only a loose simulation of real communication. Thus, we cannot rule out that participants use novel strategies when taking part in an experiment that they would not use when in an ordinary conversation.

Without additional data collection, such as eye-tracking, we also cannot be certain exactly how speakers divide their attention between target and competitor utterances during any particular trial. This disadvantage appears to become especially relevant when the paradigm is modified to allow for speeded word production, as in Experiment 6 below.

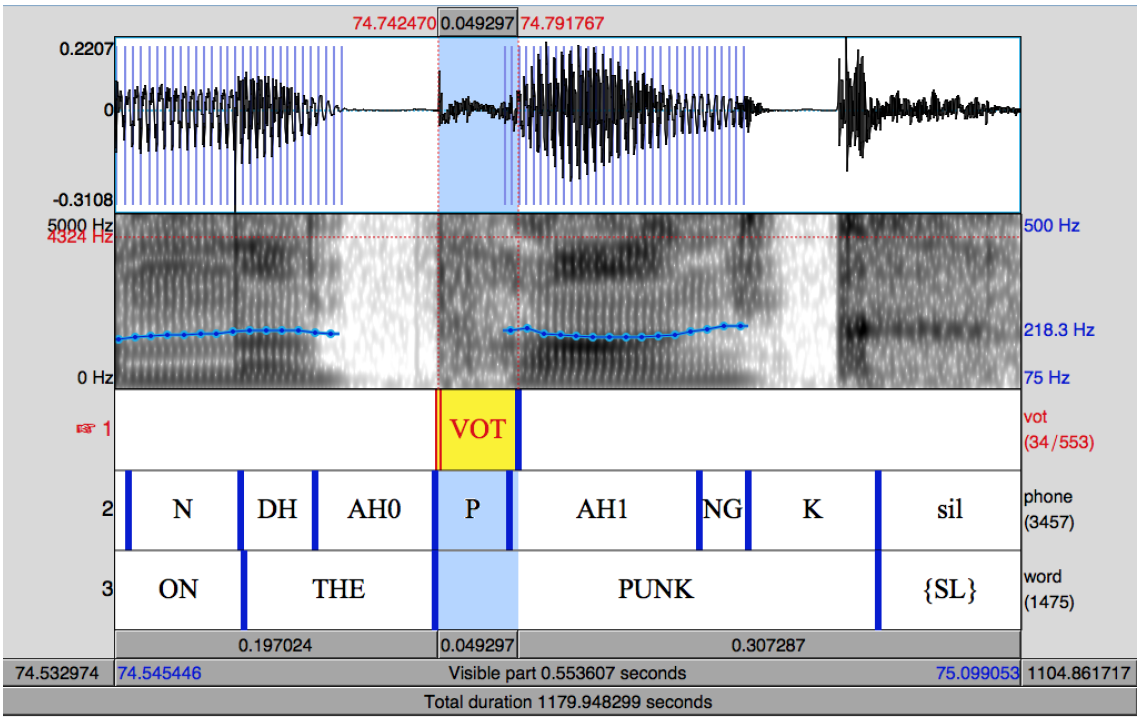

Figure 3: Measurement of the initial VOT of the target utterance "punk" using the Praat software. An automated first pass of the data finds the approximate locations of words and phonemes in the acoustic signal, as well an initial guess for the VOT. This guess is then hand-corrected.

2.1.2 Primary Dependent Variable - Voice-Onset Time (VOT)

Except for Experiment 1, all of the experiments described below use voice-onset time (VOT) as a dependent measure. VOT is the period of voiceless aspiration that follows the release of a stop consonant (i.e. /p/, /t/, /k/, /b/, /d/, /g/). The onset of VOT is typically marked by a burst of spectral energy. VOT ends when a following vowel begins. In English, voiceless stop consonants (/p/,/t/, and /k/) are identified by their long VOT values, although /p/ typically carries significantly less aspiration than either /t/, or /k/ (Hillenbrand et al. 1995). Figure 3 shows the measurement of the VOT of the initial /p/ in the token "punk" using the Praat software.

As discussed in the introductory chapter, VOT is known to vary as a function of its phonological context context (VanDam & Silbert 2010). Different consonants and places of articulation have

different characteristic VOT distributions depending on the language (Cho & Ladefoged 1999). VOT, like other phonetic features, tends to be susceptible to strengthening at the edges of prosodic domains (Fougeron & Keating 1997). VOT is also known to vary significantly as a function of global factors like speech rate, although these effects vary significantly across speakers (Theodore et al. 2007).

As mentioned in the introductory chapter, VOT has also been shown to be sensitive to competitive factors. Words from high-density phonological neighborhoods tend to be expressed with longer initial VOT (Goldinger & Summers 1989). Similarly, the Baese-Berke & Goldrick (2009) study on which this chapter is methodologically based showed that the VOT of voiceless-initial words lengthened in the context of voiced competitors.

2.1.3 Experiment 1

The introductory chapter of this dissertation introduced the idea that increased vowel space dispersion — average Euclidean distance from the center of the speakers's vowel space (Bradlow et al. 1996) — was a measure of vowel hyperarticulation that would be present under conditions of higher competition. Munson & Solomon (2004), for example, found increased vowel space dispersion in the isolated productions of words with large lexical neighborhoods. If competition affects vowels in a similar way as it did VOT in Baese-Berke & Goldrick's (2009) study, we would expect them to be hyperarticulated via increased dispersion when presented in the context of a neighbor. That is, the effects of competition could be said to be uniform across different segment types — competition would induce measurable hyperarticulation appropriate for each type. If not, some segment types may be more resilient to competitive effects.

Experiment 1 examined whether this was true. We presented target words in the context of competitors that differed from the target in vowel position, and looked for hyperarticulation in the

target vowels in the form of vowel-space expansion. Vowel space expansion has previously been used as an indicator of vowel hyperarticulation in various studies of offline effects (e.g., Munson & Solomon 2004, Wright 2003). In those studies, which were based on a simple list reading task, vowel space expansion was shown to increase as the target word's lexical frequency decreased and its neighborhood density increased.

In the present study, a condition where the targets were presented with unrelated filler words was used as a baseline for comparison. Table 1 summarizes these conditions. In addition, each target word that had a vowel-differing neighbor in the lexicon (e.g., *calm* and *comb*) was paired with another target word that had no vowel-differing neighbor (*palm* but not e.g., *pilm*). Words without competitors were presented only with fillers. This allowed the experiment to mirror a comparison performed by Baese-Berke & Goldrick between the trials on which target words with lexical neighbors were presented with unrelated fillers to those in which target words without lexical neighbors were presented with fillers. Baese-Berke & Goldrick found that the former had longer VOT, suggesting that the presence or absence of certain neighbor types in the lexicon influences online hyperarticulation effects.

Table 1: Table of conditions for Experiment 1.

Target	Vowel	Filler
CAT	KIT	DOLL

Participants

Eighteen adult native English speakers participated in this study. The participants were all undergraduate students at Johns Hopkins University who received either course credit or $10 for their involvement. The study was approved by and performed in accordance with the regulations of the Johns Hopkins University Institutional Review Board.

Materials

The stimuli used in the experiment were monosyllabic, monomorphemic, CVC content words drawn from the English Lexicon Project (Balota et al. 2007) (unless otherwise stated, all data used throughout this chapter was taken from this source). All words had a familiarity rating of 6 or above, so each word was likely to be easily recognized by each participant (Balota et al. 2007) There were 16 pairs of target words. One member of each pair had at least one neighbor in the lexicon that differed from it only in the V position; the other member of the pair had no such neighbor. The pairs were matched via paired t-test to have similar lexical frequency ($p > 0.70$) and relative frequency-weighted neighborhood density ($p > 0.13$) and phonotactic probability ($p > 0.20$). Members of each pair were constrained to have the same rhyme. Table 24 shows the set of target words used, along with their onscreen competitors for each condition.

The frequency measure used for this and all subsequent experiments was the word-form frequency taken from the Hyperspace Analogue to Language (HAL) corpus (Lund & Burgess 1996). Relative density was calculated as in Scarborough (2004). It is the log frequency of the target word divided by the sum of the word's log frequency and log frequency of all of the word's competitors. competitors and their frequencies were obtained from the English Lexicon Project (Balota et al. 2007). Phonotactic probability was calculated using the sum bigram measure from the Vitevitch phonotactic calculator at http://people.ku.edu/ mvitevit/PhonoProbHome.html (Vitevitch & Luce 2004).

Procedure

The experiment was performed using a modification of the standard paradigm described above and schematized in Figure 2. In each trial, only two words appeared on the screen of the speaker and the listener. One of the words was the target to be spoken by the speaker and subsequently identified by

the listener. The remaining word was either a filler word unrelated to the target or a lexical neighbor of the target. Babble-like noise was present in the headphones of the speaker and listener for every trial of experiment, as an incentive for the speaker to speak clearly.

Each run of the experiment consisted of 64 trials, presented in random order to each participant. There were four types of condition evenly distributed across the experiment. In the On-screen Competitor condition, the target word was presented with a vowel neighbor. In the Off-screen Competitor condition, the same target words were presented with an unrelated filler in place of the neighbor. The No Competitor condition contained target words that do not have lexical vowel neighbors; these targets were presented with unrelated fillers. Finally, Filler-only trials consisted of words unrelated to the experimental purposes paired with other, unrelated words. No data was extracted from Filler trials, and they do not factor into any experimental analyses.

To avoid confounds with potential repetition effects on the targets with lexical vowel neighbors, which appear in both the On-screen and Off-screen competitor conditions, the condition for each of these words was counter-balanced across participants. The 16 words were arbitrarily split into two sets of 8 words each. One half of the participants saw the first set of 8 words in the On-screen Competitor condition first, and the second set in the Off-screen Competitor condition second. For the remaining participants, the order was reversed. They saw the second set of 8 words in the context condition first, and the first set of 8 words in the no context condition first. On-screen target position was also counter-balanced across participants.

Acoustic and Statistical Analysis

The dependent acoustic variable measured was the Euclidean distance (displacement) of each target vowel from the center of each participant's vowel space (defined as the participant's mean F1 and F2 formant values). The boundaries of participants' vowels were manually marked using Praat, and

F1 and F2 values were extracted from the midpoint of each vowel (Bradlow et al. 1996).

The data was analyzed with linear mixed effects regression using the MCMCglmm package in R (Hadfield 2010).[15][16] A fixed effect was included for the experimental condition. Two separate models were constructed, one in which the On-screen and Off-screen competitor conditions were compared, and one in which the Off-screen and No Competitor conditions were compared. Both models used treatment coding for the fixed effect, with the Off-screen and No Competitor conditions serving as baseline conditions for the two models, respectively. Baselines were chosen as they were expected to show a lower amount of hyperarticulation within their respective models. In both models, random effects were included for participants and target items, with random slopes for condition included in each random effect as is standard in mixed effects modeling.

Results and Discussion

Results of two pairwise comparisons are summarized in Table 2. There was no significant difference in vowel displacement between the On-screen and Off-screen Competitor conditions. For words that have at least one vowel-differing neighbor, the onscreen presence of that neighbor did not induce a significant effect on vowel pronunciation in this experiment. This result is unlike what Baese-Berke & Goldrick found with respect to VOT lengthening in the context of minimal-pair competitors differing in the voice of the initial consonant.

In addition, there was no significant difference between the Off-screen Competitor and No Competitor conditions. Words that have a vowel neighbor in the lexicon did not seem to behave differ-

[15]MCMCglmm was run for 20,000 iterations. All other parameters (burn-in, thinning) were left at their default values.

[16]The commonly used lme4 R package does not support the calculation of p-values for models which include random slope parameters, which rendered it less suitable for the analyses in this dissertation. lme4 only supports approximate p-values for models without random slopes via Parameter Markov Chain Monte Carlo (pMCMC) (Baayen 2008).

ently from words that do not. This result is also at odds with what Baese-Berke & Goldrick (2009) found with respect to VOT.

Table 2: Experiment 1: Statistical results .

Comparison	Effect Mean	Lower 95% CI	Upper 95% CI	p
On-screen vs. Off-screen	-1.997	-22.915	18.786	< 0.8
Off-screen vs. No Competitor	1.408	-153.347	155.551	< 1.0

As the present study closely followed the setup of used by Baese-Berke & Goldrick (2009), the apparent lack of significant effects is surprising, although another study also using a variant of the Baese-Berke & Goldrick paradigm also failed to find increased vowel-space dispersion (Lefkowitz 2012) . There are several possible explanations for why this might be the case. There were minor methodological differences between the present study and the Baese-Berke & Goldrick (2009) study. The Baese-Berke & Goldrick study did not include noise and each trial presented the participants with three words, a target and either two fillers or a neighbor and filler. In the present experiment, every trial was presented with noise and there were only two words on screen at a time. It seems unlikely that these differences would cause such a disagreement in results, unlike the substantive effects discussed in the text below. Nevertheless, subsequent experiments described in this chapter hewed more closely to B&G's procedures in order to decrease the chance of results varying for methodological reasons.

While dispersion is widely used as a metric for vowel hyperarticulation (Munson & Solomon 2004), actually measuring it poses some challenges compared to measuring VOT. It is possible to measure the VOT of each target item independently of the other target items, but the displacements used as the dependent variables in the present experiment were calculated relative to the mean F1/F2 value of all of the targets in a particular condition. As this mean depends on several targets, the displacement measures of these targets are not truly independent. If vowel space-expansion effects are very small relative to the overall variability of vowel production, this lack of independence

would exacerbate a lack of experimental power.

In addition, if hyperarticulation is defined as more effortful production that tends to differentiate an utterance from similar utterances, then outward displacement from the mean of a speaker's vowel space may not always represent a hyperarticulated production. For example, a CVC word can have several lexical competitors that differ in only the vowel. A similar stop-initial word can only have one lexical neighbor that differs in the VOT of its initial segment (e.g., /p/ vs./b/ or /k/ vs./g/), though it can have many competitors that differ in other features such as place of articulation. This means that it is not as clear how a particular vowel should be displaced to differentiate itself from the vowels in a target's competitors. In some cases, bringing a vowel closer to the participant's vowel center may serve to differentiate it better than moving it farther away. Indeed, Neel (2008) recently showed that vowel space expansion itself was not as good an indicator of intelligibility as the individual distances between nearby pairs of vowels. Finally, Gahl et al. (2012) suggest that vowel space expansion, despite its widespread use, may not actually be a valid indicator of competition-induced phonetic variation at all. Their study showed that the offline vowel space expansion effects found by Munson and Solomon (2004) and Wright (2003) were almost entirely predicted by the phonological context of the target vowels (the features of the preceding and following consonants). However, when Lefkowitz (2012) explicitly attempted to use 'disassociation' (distance from the nearest vowel) as a measure of vowel hyperarticulation rather than dispersion in a Baese-Berke & Goldrick (2009) style experiment, he also failed to find any significant effects.

Aside from potential measurement pitfalls, there may be theoretical implications of the fact that there was a difference in VOT between the On-screen and Off-screen competitor conditions in the Baese-Berke & Goldrick study, but not a similar difference in vowel displacement in the present study. It may be when a word is presented on-screen with its neighbor, competition between them, although due to a difference in the vowel position, is at the lexical whole-word level of processing. If

that is the case, speakers may use a generic hyperarticulatory strategy that enhances the target word as a whole rather than just the competing vowel. One strategy the speaker may use is to amplify the word's initial segments. This would be consistent with previous studies showing the importance of initial segments in both production and perception. Fougeron et al. (1997) showed that articulatory strengthening occurs at the edges of prosodic domains (the initial segments of a word are often at these edges). Marslen-Wilson et al. (1989) showed that initial segments were more important for lexical access than other parts of a word. Vitevitch (2002) showed that, all other things being equal, words with more onset competitors (competitors in the lexicon sharing the same initial segments as the target word) were associated with slower and less accurate recognition. If competition was holistic in this way, then we might expect that on-screen vowel-differing competitors could induce hyperarticulation in the onsets of target utterances. This possibility is explored in Experiment 2 below.

By itself, however, such a mechanism would not be sufficient to explain the VOT difference found by Baese-Berke & Goldrick between the Off-screen and No Competitor conditions, as there was no on-screen competitor present in either. One possibility is that the increased VOT found in the Off-screen condition was not due to the existence of a voiced minimal pair neighbor in the lexicon, but due to larger overall neighborhood density for the words in the Off-screen condition as opposed to those in the No competitor condition. The words in these two conditions were not explicitly matched on neighborhood density in the Baese-Berke et al experiment. In fact, a two-sample paired t-test comparing the average frequency-weighted neighborhood densities[17] of the two sets of words used in the Baese-Berke experiment shows a significant difference ($p < 0.05$).

[17] As calculated from the English Lexicon Project using the procedure in Scarborough (2004).

Words in the Off-screen condition have higher average neighborhood density than no-competitor items. By comparison, the same test performed on the words used in the present experiment shows that the two sets of words are matched on neighbor density ($p > 0.13$).

If the difference found by Baese-Berke & Goldrick was indeed due to differences in aggregate lexical factors rather than the presence or absence of particular neighbor types in a word's neighborhood, their result would be consistent with previous studies suggesting offline hyperarticulation effects do not depend on the phonological structure of a word's neighborhood (Scarborough 2004; Goldinger & Summers 1989).

2.1.4 Experiment 2

Baese-Berke & Goldrick found VOT lengthening in the initial segments of target words presented with an onscreen neighbor that differs in the voicing of its initial segment. Experiment 2 was designed to determine if this VOT lengthening is due to competition at a whole-word lexical level of processing, or post-lexical position-level processing (i.e., are the phonetic properties in a particular position of the target word enhanced only in the presence of a competitor that differs at that position)? Again, if the relevant competition is at the lexical level, the effects of a competitor should be visible across the entire target form. On the other hand, if the relevant competition was at the post-lexical level, the effects of a competitor would only be visible in those positions in the target that differ from the competitor.

Target words were presented in the context of competitors that differed in onset (a replication of the Baese-Berke & Goldrick study using voice-differing competitors), vowel, and coda positions. They were also presented with only unrelated filler words as a baseline. The four conditions under which VOT was measured are summarized in Table 3. Tables 25, 26, and 27 show the full set of target words used, along with their onscreen competitors for each condition.

Table 3: Table of conditions for Experiment 2.

Target	Onset (Voice)	Vowel	Coda	Filler A	Filler B
CAP	GAP	CUP	CAT	WOLF	DIM

Participants

Twenty-four adult native English speakers participated in this study. The participants were all undergraduate students at Johns Hopkins University who received either course credit or $10 for their involvement. The study was approved by and performed in accordance with the regulations of the Johns Hopkins University Institutional Review Board.

Materials

The materials for this experiment were 48 monosyllabic CVC target words. Each target began with a voiceless stop (16 with /p/, 19 with /t/ and 13 with /k/). As in Baese-Berke & Goldrick (2009), some of the words were proper names. For each target word, three competitor words were chosen that had spellings similar to the target. One of the competitors differed from the target in the voicing of the initial consonant, one differed in the vowel, and one differed in the coda. In addition, two phonetically unrelated words were chosen for each target word to serve as fillers. The stimuli used are shown in Tables 25, 26, and 27. Different competitor types were matched for lexical frequency using paired t-tests (Onset vs. Vowel Competitors $p > 0.6$, Onset vs. Coda Competitors $p > 0.3$, Vowel vs. Coda Competitors $p > 0.4$).

Procedure

The procedure was similar to that used in Experiment 1. However, the stimuli were presented without noise and three (instead of two) words were displayed on each trial. This was done to better match the procedure used by Baese-Berke and Goldrick. On critical trials, the target word was

presented with either a competitor and an unrelated filler, or two filler words. As in Experiment 1, all-filler trials were also presented but not analyzed. All words were pronounced within the carrier phrase "Click on the...".

To create the trials, 4 lists were created. Each target word was pseudo-randomly assigned to one of the four measured conditions in each list, with the constraint that the number of target words of each onset type (/p/,/t/,/k/) was approximately equal in each condition in each list. Each participant was presented with only one of the lists, and so only saw each target in one of the 4 measured conditions. Every 4 participants, each word appeared once in each of the 4 measured conditions. Each participant was not exposed to every trial type for every target in order to avoid strong repetition effects (repeated productions of the same word within a short time frame tend to be significantly reduced (Fowler 1988; Lehnert-LeHouillier 2010)) and to keep experiment length manageable.

Each of the 4 lists of 48 trials was augmented with 48 filler trials, making 4 sets of 96 trials. The order of the trials within a list was randomized by participant. On-screen target position was counterbalanced across participants. In order to balance the need to collect data from as many trials as possible with the need to avoid the repetition effects mentioned above, each participant ran through the list twice, for a total of 192 trials per run of the experiment. The list order was the same in both runs, as this ensured that the time between repetitions of the same word remained uniformly long.

Acoustic and Statistical Analysis

The VOTs of the target words were manually measured using Praat. The data was analyzed with linear mixed-effects regression using the MCMCglmm package in R. The model used treatment coding for the fixed effect of competitor type (Onset, Vowel, or Coda), with the Filler condition as a baseline. Random effects were included for participants and target items, with random slopes for

competitor type included in each random effect.

Results and Discussion

Results of a linear mixed results regression comparing each neighbor type to a filler baseline are shown in Figure 4 and Table 4. Errors bars in Figure 4, and all subsequent bar plots in this chapter, represent the standard error of the mean ($\pm\frac{\sigma}{\sqrt{N}}$) across all N utterances in a particular condition. Only onset-differing competitors appear to cause a significant VOT enhancement effect over fillers. This result suggests that the mechanisms that cause VOT lengthening involve processing that is position-specific. This means that for initial VOT to lengthen, it must be in the context of a competitor that differs in initial position. This might imply that other parts of the target word could be hyperarticulated in the context of competitors that differ in the those positions. However, the null results in Experiment 1 suggest that this is not necessarily the case. Vowel-differing competitors do not seem to induce hyperarticulation in either the vowel *or* onset positions of targets. One possibility is that initial positions are privileged during processing so that differences in other positions are less salient, and thus have less of an effect on phonetic realization. If this were true, it might be difficult to find specific enhancement effects anywhere but word-initially.

Table 4: Experiment 2: Statistical results.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Onset (Voice) Competitor	1.96614	0.34569	3.61954	< 0.02*
Vowel Competitor	-0.03148	-1.67714	1.51457	< 1.0
Coda Competitor	0.52178	-1.02070	2.37164	< 0.6

Interestingly, the effects found seem to be limited mainly to the first production of each target word. Second productions show no significant VOT difference across conditions, suggesting a strong effect of repetition. Furthermore, as shown in Figure 5, the effects found are limited to cases when the target word begins with /p/ or /t/. This may be due to a ceiling effect associated with the

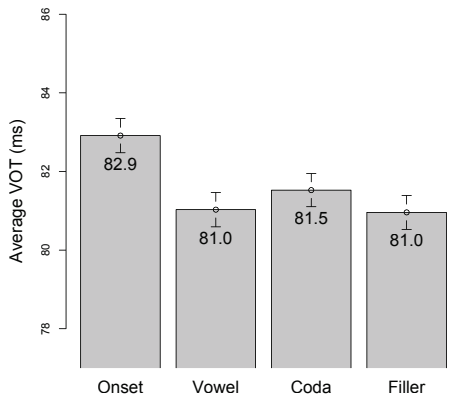

Figure 4: Experiment 2: Comparison of mean VOT across experimental conditions.

/k/-initial targets used in the experiment, as /k/-initial words are known to have long base VOTs that participants may not be able to lengthen further.

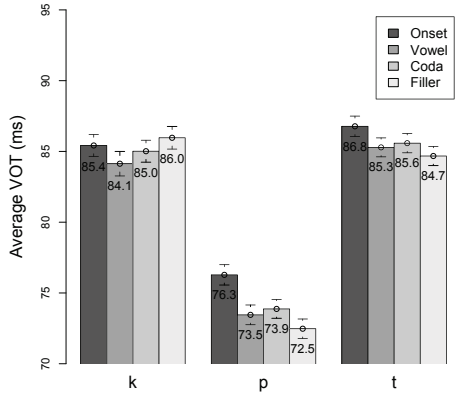

Figure 5: Experiment 2: VOT broken down by target onset phoneme and condition.

2.1.5 Experiment 3

The results of Experiment 2 suggest that online hyperarticulation effects are 'local'. That is, they are likely occurring at a post-lexical positional level of processing. The goal of this experiment, which

consists of two sub-experiments, was to determine if the similarity metric that determines whether a competitor induces online VOT enhancement is broad or narrow. In particular, we tested to see if only certain kinds of onset competitors induced VOT enhancement. The experimental procedure and data preparation in both sub-experiments was identical to that used in Experiment 2.

The first sub-experiment (referred to as Experiment 3A) was designed to look for an enhancement effect in the context of place-of-articulation differing onset competitors. The set of experimental conditions is shown in Table 5. All of the place competitors in Experiment 3A differ from the target in just one place of articulation (place) feature.

Table 5: Table of conditions for Experiment 3A.

Target	Voice	Place	Filler A	Filler B
CAPE	GAPE	TAPE	NUN	SHED

The second sub-experiment (referred to as Experiment 3B) was designed to look for an effect of competitors differing from the target only in the manner of their initial consonant. Competitors were chosen so as to differ minimally from the targets with respect to manner, but compromise was necessary due to the limitations of the phoneme inventory of English. In particular, /p/-initial targets were paired with /f/-initial competitors, which differ in manner and a minor place feature (labial vs. labiodental) and /t/-initial targets were paired with /s/-initial competitors, which differ in manner and stridency. Unfortunately, English does not use the velar fricative /x/. As a compromise, /k/-initial competitors were paired with /h/-initial competitors, which differ in manner and place. The set of experimental conditions is shown in Table 6.

Table 6: Table of conditions for Experiment 3B.

Target	Voice	Manner	Filler A	Filler B
PUN	BUN	FUN	LARD	SHIP
TEEM	DEEM	SEEM	WET	LOUD
KILT	GUILT	HILT	TOOL	VENT

Participants

Twenty-two adult native English speakers participated in this study. The participants were all undergraduate students at Johns Hopkins University who received either course credit or $10 for their involvement. The study was approved by and performed in accordance with the regulations of the Johns Hopkins University Institutional Review Board.

Materials

Experiment 3A used 33 monosyllabic CVC target words (11 /p/-initial, 11 /t/-initial and 11 /k/-initial). The full set of stimuli used, including competitors and fillers for the targets, are shown in Tables 28, 29, and 30. Competitor types were matched for frequency (paired t-test, $p > 0.8$).

Experiment 3B used 36 monosyllabic CVC target words (12 /p/-initial, 12 /t/-initial and 12 /k/-initial). The full set of stimuli are shown in Tables 31, 32, and 33. Again, competitor types were matched for frequency (paired t-test, $p > 0.8$).

Procedure

The procedure was identical to that used in Experiment 2, except for the following. Each run of Experiment 3A included 132 total trials (participants ran through a list of 33 target trials and 33 fillers trials twice). A run of Experiment 3B included 144 total total trials. Each of the twenty-two participants that took part in the study was run through both Experiment 3A and 3B, although half the participants took part in Experiment 3A first, and half took part in Experiment 3B first. Thus, each participant saw a total of 276 trials.

Acoustic and Statistical Analysis

As in Experiment 2, target VOTs were measured using Praat, and analyzed using MCMCglmm. The setup of the statistical model used was the same as in Experiment 2. In the model, data from both Experiment 3A and 3B were joined and analyzed together. So, the fixed condition was coded for competitor type — Voice (from experiments 3A and 3B), Place (from experiment 3A), and Manner (from experiment 3B) — and used the Filler condition (from experiments 3A and 3B) as a baseline.

Results and Discussion

Results for Experiment 3A only are shown in Figures 6 and 7. Results for Experiment 3B only are shown in Figures 8 and 9. Statistical results for Experiment 3 as a whole are shown in Table 7.

There was a significant VOT enhancement effect of place competitors, and the effect is consistent across /p/,/t/, and /k/-initial targets. It is interesting that the VOT of /p/ lengthens in the context of /k/ and /t/, given that /k/ and /t/ tend to have longer average VOT than /p/, and thus VOT lengthening might make /p/ initial words more similar to their competitors (Cho & Ladefoged 1999). However, aspiration contains other cues besides its length that signal place of articulation, and increasing VOT may strengthen these alternate cues (Suchato & Punyabukkana 2005; Repp & Lin 1988).

There appears to be no overall significant effect of manner competitors on VOT enhancement. However, the breakdown of the results by target onset (Figure 9) indicates that there is a potential enhancement effect for /p/ onsets in the context of /f/ initial competitors. Statistics for just the /p/-initial targets in Experiment 3 are shown in Table 8. Since /p/ is likely more similar to /f/ than /k/ is

to /h/ (differing in a major place feature) or /t/ is to /s/ (differing in stridency)[18], it may be that online VOT enhancement may only occur in the context of competitors that are sufficiently similar to the target word — about one phonetic/phonological feature away. Thus, competition is determined by a fairly narrow similarity metric. It is also important to note that while /k/ and /t/ initial targets did not show VOT lengthening in the context of manner-differing competitors, they still showed lengthening in the context of voice-differing competitors — a replication of previous results.

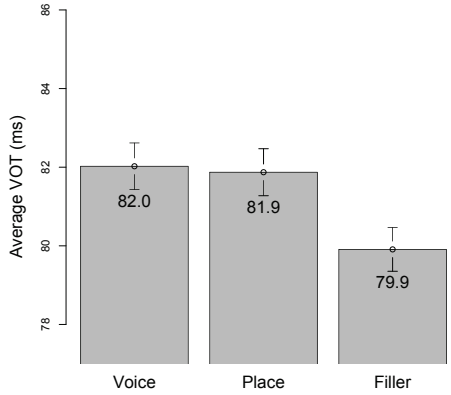

Figure 6: Experiment 3A: Comparison of mean VOT across experimental conditions.

Table 7: Experiment 3: Statistical results.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Voice Competitor	2.7109	0.7963	4.6534	< 0.002*
Place Competitor	3.1722	1.0126	5.3568	< 0.02*
Manner Competitor	0.2196	-1.9156	2.2156	< 0.9

[18]Although /p/ and /f/ are more likely to be confused perceptually than /t/ and /s/ or /k/ and /h/, according to experimentally-derived confusion matrices (Benkí 2003; Luce 1986), the effects found might not be due to their apparent similarity; instead, they may simply be a property of /p/-initial targets. It is difficult to disentangle this question using English stimuli, since another stop/fricative pair as similar as /p/ and /f/ does not exist in the phoneme inventory.

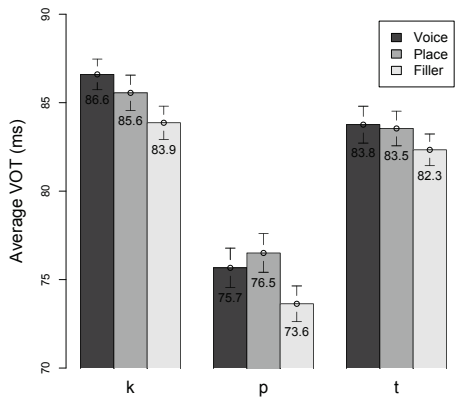

Figure 7: Experiment 3A: VOT broken down by target onset phoneme and condition.

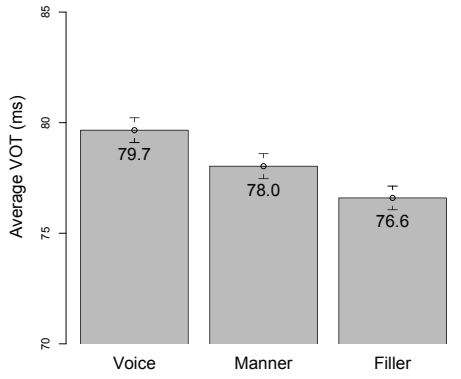

Figure 8: Experiment 3B: Comparison of mean VOT across experimental conditions.

2.1.6 Experiment 4

If the interpretation of Experiments 2 and 3 as presented above is correct and it is indeed the case that contextual competitors must be sufficiently similar (about one phonological feature difference) to target words to induce hyperarticulation, then we would predict that more distant competitors would not have any significant effect. Experiment 4 aims to support this prediction with the use of nasal-initial competitors for /t/ and /p/ initial targets. Unfortunately, English lacks the sound /ŋ/ word-

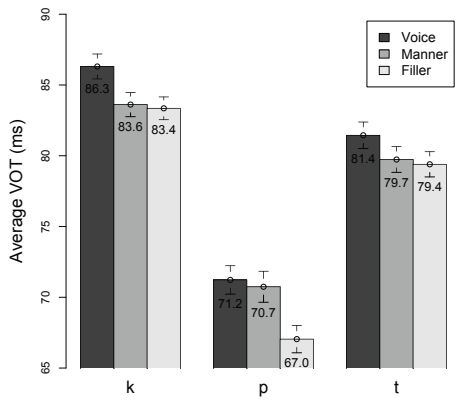

Figure 9: Experiment 3B: VOT broken down by target onset phoneme and condition.

Table 8: Experiment 3: Statistical results for /p/-initial targets.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Voice Competitor	3.3835	0.7123	5.8069	< 0.002*
Place Competitor	4.8233	1.6259	7.6235	< 0.002*
Manner Competitor	1.6583	-1.2322	4.7638	< 0.3

initially. As a substitute, I used null-onset competitors for /k/ initial targets. English phonology is ensures that words that begin with a vowel underlyingly surface with an initial glottal stop /ʔ/. Both nasal and null-initial competitors differ significantly from their voiceless stop counterparts in terms of abstract phonological features and perceptual similarity (Benkí 2003; Bailey & Hahn 2005). Nasals differ from voiceless stops in both nasality and voicing. Glottal stops differ from /k/ in both place and voicing.

Table 9: Table of conditions for Experiment 4.

Target	Voice	Nasal/Null	Filler A	Filler B
PILL	BILL	MILL	HAIR	FOOD
TAME	DAME	NAME	BENCH	SIGN
KILL	GILL	ILL	REED	NOON

Participants

Twenty-four adult native English speakers participated in this study. Data from one participant was excluded as the distribution of their VOT values was significantly different form the expected English distribution (Hillenbrand et al. 1995; Lisker & Abramson 1970). The participants were all undergraduate students at Johns Hopkins University who received either course credit or $10 for their involvement. The study was approved by and performed in accordance with the regulations of the Johns Hopkins University Institutional Review Board.

Materials

This experiment used 51 monosyllabic CVC target words, including some proper names (17 /p/-initial, 16 /t/-initial and 18 /k/-initial). Targets, along with their competitors and fillers are shown in Tables 34, 35, and 36. As usual, competitor types were matched for frequency (paired t-test, $p > 0.8$).

Procedure

Experimental procedure was identical to that used for Experiment 2. Each run of the experiment included 204 trials (a list of of 51 competitor trials and 51 filler trials was run through twice by each participant).

Acoustic and Statistical Analysis

Data collection and analysis were identical to that used in Experiment 2.

Results and Discussion

Results are shown in Figures 10 and 11 and Table 10. As predicted, there appears to be no overall significant effect of nasal competitors or glottal-stop initial competitors on VOT length. The effects of the voiced onset competitor are numerically in the expected direction and around significance ($p < 0.05$) across multiple runs of MCMCglmm, at least partially replicating the results of previous experiments. The lack of a strong significant result in this experiment is difficult to explain. It likely isn't due to a lack of sufficient data points, since the number of items used was comparable to Experiment 2 (although significantly less than Experiment 3). It is also unlikely to be caused by the particular properties of the target set, as their is substantial overlap with the targets used in the previous experiments.

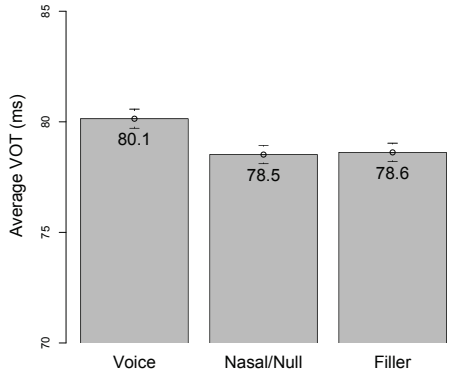

Figure 10: Experiment 4: Comparison of mean VOT across experimental conditions.

Table 10: Experiment 4: Statistical results.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Voice Competitor	1.544356	0.007767	3.317701	< 0.05*
Nasal/Null Competitor	-0.104006	-1.724541	1.547358	< 0.9

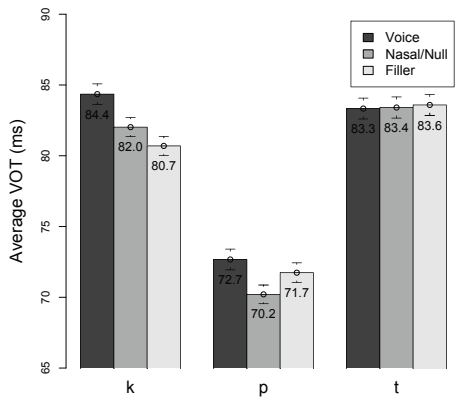

Figure 11: Experiment 4: VOT broken down by target onset phoneme and condition.

2.1.7 Experiment 5

If it is the case that the level of similarity between targets and competitors is important for inducing hyperarticulation, then it may be that the null results found in Experiments 1 and 2 with respect to the effect of vowel competitors on vowel space expansion and initial VOT, and coda competitors on initial VOT, are the result of the similarity of these competitors to the target words not being carefully controlled. For example, in Experiment 2, PUN could serve as a coda-differing competitor for PUB, even though /n/ and /b/ are significantly different sounds. Thus, the apparent positional specificity suggested by the results of Experiments 1 and 2 might be an artifact of the experiments not including sufficiently similar vowel and coda competitors.

Experiment 5 replicates the coda condition of Experiment 2 in an attempt to determine if initial VOT is enhanced as a result of competition with sufficiently similar coda-differing competitors. To this end, all coda-competitors in Experiment 5 different from the target utterance in exactly one feature, either place or voicing. Experiments 2 and 3 indicate that place and voicing differences do cause VOT enhancement when in initial position. Table 11 summarizes the conditions used in

Experiment 5.

Table 11: Table of conditions for Experiment 5.

Target	Voice/Place	Coda	Filler A	Filler B
PEEP	KEEP	PEAT	DART	FOG
TUG	DUG	TUCK	MACE	YAM
COAT	GOAT	CODE	LACE	BEAN

Participants

Twenty-five adult native English speakers participated in this study. Data from one participant was excluded due to equipment failure. The participants were all undergraduate students at Johns Hopkins University who received either course credit or $10 for their involvement. The study was approved by and performed in accordance with the regulations of the Johns Hopkins University Institutional Review Board.

Materials

The materials for this experiment were 42 monosyllabic CVC target words (15 /p/-initial, 12 /t/-initial and 15 /k/-initial). Targets, competitors, and fillers are shown in Tables 37, 38, and 39. For each target word, both competitors differed in the same feature (if the onset competitor differed in voicing, so did the coda competitor). Onset and coda competitor types were matched for frequency (paired t-test, $p > 0.8$).

Procedure

Experimental procedure was identical to that used for Experiment 2. Each run of the experiment included 168 total trials.

Acoustic and Statistical Analysis

VOTs were measured as in Experiment 2. An MCMCglmm model of the measurements was created with an identical structure to that used in Experiment 2. As usual, the model used the Filler condition as a baseline for the fixed effect. Four competitor conditions were created for comparison to the baseline by crossing the position-type of the competitor (Onset, Coda) with the feature-type of the competitor (Voicing, Place).

Results and Discussion

Results are shown in Figures 12 and 13, and Table 12. The results indicate that even when competitor codas are controlled to be one phonological feature away from target codas, there is no significant VOT enhancement effect induced. This provides more evidence that, as suggested in Experiments 1 and 2, the speech planning mechanisms that respond to competition and produce VOT enhancement are position-specific. Similarity in the codas of targets and competitors does not appear to significantly affect the phonetic realizations of word onsets, while similarity in the onsets does.

According to the data, the only potential VOT lengthening above the baseline filler condition occurs when the target is presented with a competitor differing in initial voicing. The effect does not reach significance, but this might be a power issue, as the number of trials in which voiced onset competitors were presented was limited. Onset competitors differing in place of articulation did not appear to have an effect in this experiment. This was unexpected given the results of Experiment 3 above, but it is not the first time place effects have proven difficult to replicate. Schertz (2013) also could not find an effect of onset place competitors on target productions in a task that asked participants to correct target productions that were misheard as competitors. She did, however, find the expected VOT-lengthening effect of voiced onset competitors.

It is possibly relevant to note that, due to issues of experimental design, the target set in Experiment 3 and target set in the current experiment are almost entirely disjoint. Many of the target words in Experiment 3 end in sonorants (i.e., /n/,/m/,/r/,/l/), while all of the targets in the current experiment end in stop consonants (often voiceless) in order to ensure that they can be paired with an appropriate coda competitor. This difference in the phonological structure of the targets might be responsible for the lack of an effect of place-differing neighbors in the current experiment. It is well-known that vowels before voiceless consonants are shorter than in other contexts. This shortening may also be manifested in the VOT of the pre-vowel consonant. Port & Rotunno (1979) found a correlation between VOT and vowel length — VOT tended to be shorter before shorter vowels (e.g., lax vowels in English). Languages also tend to avoid grouping similar sound together — an effect known as the obligatory contour principle, or OCP. In this case, having two voiceless sounds in close proximity may result in shortening of the first consonant in order to differentiate it from the second.

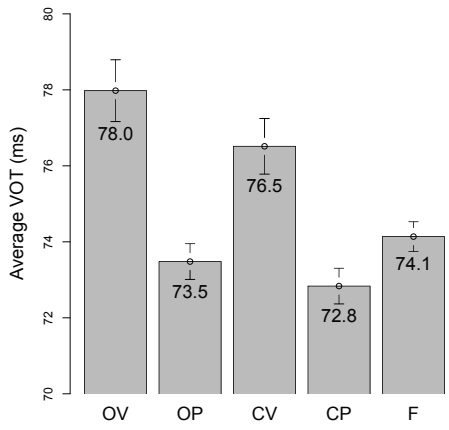

Figure 12: Experiment 5: Comparison of mean VOT across experimental conditions.

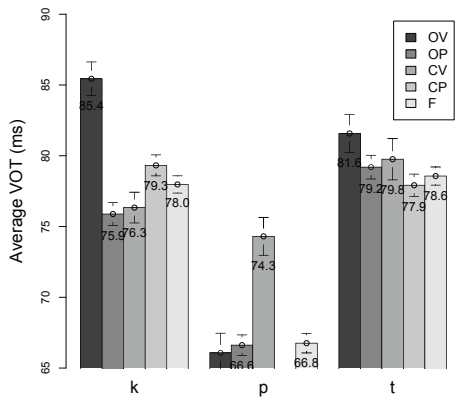

Figure 13: Experiment 5: VOT broken down by target onset phoneme and condition.

Table 12: Experiment 5: Statistical results.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Onset Voice Competitor	1.9332	-0.9687	4.5215	< 0.2
Onset Place Competitor	0.1339	-1.5087	1.9723	< 0.9
Coda Voice Competitor	-0.2883	-2.7891	2.3830	< 0.9
Coda Place Competitor	-0.1929	-1.9612	1.3744	< 0.9

2.1.8 Summary of Hyperarticulation Experiments

Using the Baese-Berke & Goldrick (2009) paradigm, I performed a battery of experiments to determine in what ways on-screen competitors could differ from the target utterance and induce VOT hyperarticulation. The results are summarized in Table 13. Checkmarks in the table of results indicate that at least one significant effect was found. It should be noted that results were not entirely consistent across experiments. Due to small effect sizes (usually less than 5ms) and high variability across speakers, statistical analyses sometimes ran into power issues when sample sizes were small. This is supported by the fact that significance levels were highest in Experiment 3, whose analysis aggregated data from two sub-experiments and thus include many more trials than the analyses of the other experiments. Furthermore, the results of Experiment 5 suggest that the phonological struc-

ture of the target set (i.e., the relationship between target onsets and codas) may modulate effects in an unknown way. Nevertheless, taken as a whole, the results point to the following generalization. Competition induces hyperarticulation only when competitors are sufficiently similar to the target word (a difference of roughly one phonological feature, such as voicing or place of articulation). The effect drops off quickly as phonological and/or phonetic distance increases.

Table 13: Summary of Hyperarticulation Results

Target	Competitor	Difference	Effect
CAP	DOLL	**unrelated**	X
CAP	CAD	**coda**	X
CAP	CUP	**vowel**	X
TAP	NAP	**onset voicing + nasality**	X
CAP	TAP	**onset place**	✓
CAP	GAP	**onset voicing**	✓

Hyperarticulatory effects induced by competition appear to be position-specific. For a competitor to induce initial VOT lengthening in a target, its own onset must be minimally different from that of the target (e.g., CAP and GAP). Minimal difference elsewhere does not induce VOT lengthening in the onset position (e.g., CAP and CAB). However, the set of experiments described here and in the literature are not sufficient to rule out the importance of position-independent whole-form competition for phonetic variation. Would GAP induce initial VOT lengthening in largely unrelated COT, as they share similar initial consonants, or would the competitor and target also need to be minimal pairs (e.g., GOT and COT)? While this is an empirical question that is left for future study, I will follow the intuition that GAP does not seem like a viable competitor for COT, and assume that whole-form competition is also an important factor in conditioning phonetic variation. In particular, it may serve as a 'gateway' to position-specific competition — position specific competition may only be possible when whole-word competition is more fierce. This idea is discussed further in Section 4.2 of Chapter 4.

2.2 Experiments Linking Latency and Hyperarticulation

There is an apparent nonlinear relationship between feature distance and effect size that determines the delay involved in initiating speech in the presence of competition. When targets and competitors are very similar, latency is high (e.g., Gordon & Meyer 1984). As distance increases, latency quickly drops. The hyperarticulation experiments presented above show a similar pattern. Hyperarticulation is induced only when competition exists between sufficiently similar utterances. This common pattern suggests that both types of effects may share a common cause, or at least be the result of similar mechanisms. The pattern may also be interpreted as an instance of Shephard's universal law of generalization (Shepard 1987). Specifically, the law states that the probability that one object is interpreted to be the same as another is usually an exponentially decaying function of their similarity in some appropriate psychologically defined space. In this case, the rapid drop-off of both increased latency and hyperarticulatory effects might be expected as a function of the linearly increasing perceptual or phonological similarity between them.

It has been proposed in the literature that latency and hyperarticulation are linked. Bell et al. (2009) suggest that lexical-access latency (as measured by the time it takes to initiate speech in tasks such as picture naming (Griffin & Bock 1998)) and word duration are broadly correlated. Shaw (2012) found that words with unpredictable stress patterns tended to be pronounced with both higher latency and duration. Yap (2011) found that vowel duration tended to increase, and formant values were modified, under high cognitive load tasks usually associated with higher latencies. Similarly, Munson (2013) has also found that latency in a picture naming task is a good predictor of overall vowel dispersion. To my knowledge, however, no published experiment has directly attempted to correlate response latency with VOT hyperarticulation within the same experiment. Experiment 6 was designed to test for such a correlation.

2.2.1 Experiment 6

Experiment 6 was designed to allow measurement of both how long it takes to initiate speech in the presence of competitors (a proxy measure for the time it takes to plan an utterance), and the initial VOT of the speech produced. Voiceless initial target utterances were produced with voiced initial competitors and unrelated fillers. Only voiced competitors were used, because Experiments 2-5 indicated that these are the most consistent in inducing VOT lengthening, and thus the strongest possible competitors to the target words. Table 14 summarizes the conditions used in Experiment 6. It was predicted that in high-competition situations, such as when a voiced competitor was presented on screen with the target word, participants would take longer to initiate speech, and their speech would be realized with longer initial VOTs.

Table 14: Table of conditions for Experiment 6.

Target	Voice	Filler A	Filler B
PORE	BORE	SHELF	MILE

Participants

Thirty adult native English speakers participated in this study. Data from two of the participants was excluded due to equipment failure. The participants were all undergraduate students at Johns Hopkins University who received either course credit or $10 for their involvement. The study was approved by and performed in accordance with the regulations of the Johns Hopkins University Institutional Review Board.

Materials

The experiment used 74 monosyllabic CVC target words (23 /p/-initial, 25 /t/-initial and 26 /k/-initial). Each target began with a voiceless stop (23 with /p/, 25 with /t/ and 26 with /k/). The full

set of stimuli including targets, their voiced minimal pairs, and fillers, are presented in Tables 40, 41, and 42.

Procedure

The experimental procedure used in Experiment 6 was a variant of the procedure used in Experiments 2 through 5, with some key modifications. First, participant's productions were not monitored by a listener. Once again, on every trial of the experiment, participants were initially presented with three potential target words on a computer screen, with the three words centered on a line at the vertical midpoint of the screen. After 1500ms, a square frame appeared around one of the words. Participants were given instructions to say the cued word *as quickly as possible*. On critical trials, the target word was presented with either a competitor and an unrelated filler, or two filler words. As in Experiments 2 through 5, all-filler trials were also presented but not analyzed. There was no carrier phrase, so the time between the appearance of the square cue and the initiation of speech by the participant could be directly measured. participants were given 1500ms to say the target word, at which point the experiment would automatically move on to the next trial, with a 750ms delay between trials. There were 296 total trials on each run of the experiment.

Acoustic/Chronometric Measurements

The VOTs of the target words were measured as in Experiments 2-5. Reaction times were calculated as the time between the cue to begin speaking and the onset of VOT. Manual RT measurements were used in order to avoid the systematic phonetic biases introduced by hardware voice keys (Rey et al. 2013; Kessler et al. 2002). Even with manual measurement, some phonetic biases may be unavoidable (e.g, due to the recording equipment, Praat's processing of the data, etc.). The design of the experiment mitigates this somewhat by allowing within-word comparisons across different

conditions (counterbalanced across subjects). Ideally, phonetic biases shouldn't interact with anything other than the phonological structure of target words, keeping differences due to condition constant.

Statistical Analysis of Original Conditions and Results

Basic results by condition are shown in Figures 14 and Table 15. As in Experiments 2-5, the data was analyzed with a MCMCglmm model. The random effects in the model were defined as in the previous experiments. The model used treatment coding for the fixed effect of competitor type (Voice), with the Filler condition as a baseline. Based on these overall results, it appears there is no significant effect of competitor type on participant's reaction times. There also does not appear to be an effect on VOT values, apparently contradicting Experiments 2 through 5.

Although this wouldn't account for the lack of reaction time differences, one possible factor is that VOT measurements may be less reliable in speeded speech tasks such as the one in Experiment 6 due to floor effects. The need for participants to produce responses as quickly as possible could result in a correlated maximization of speech rate. Indeed, overall VOT values in Experiment 6 are approximately 10ms lower than in Experiments 2 through 5. If there is a minimal threshold VOT for each phoneme that participants will not pass, then participants' responses may all be approximating this threshold, effectively hiding some phonetic variation that would otherwise be expected. Any measure of hyperarticulation that is duration-based (including VOT) would be susceptible to floor effects in speeded production experiments. A different measure of hyperarticulation, such as the spectral center of gravity of the burst after the closure phase of a stop consonant (Bonneau 1996; Repp & Lin 1988; Marcel et al. 1978), that is not confounded with overall speech rate could be used in future experiments attempting to correlate hyperarticulation and latency.

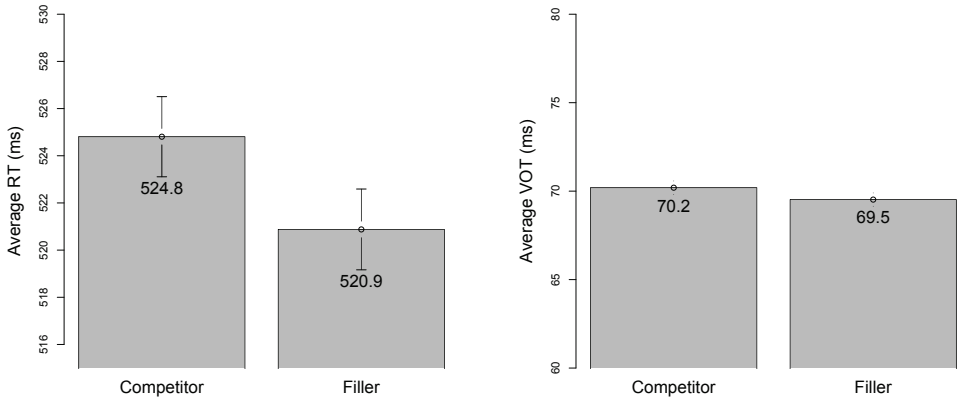

Figure 14: Experiment 6: Comparison of mean latency and VOT across experimental conditions.

Table 15: Experiment 6: Statistical results by condition.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Competitor (Reaction Time)	5.155	-1.952	11.737	< 0.2
Competitor (VOT)	0.9495	-0.4610	2.4384	< 0.2

Statistical Analysis of Recoded Conditions and Results

Despite the null results based on the original experimental conditions, a more thorough analysis of the data indicated that the position of stimuli on the screen had a large effect on reaction times. Target words presented in the center of the screen were produced 62ms faster on average than targets in the left and right positions. The apparent advantage of stimuli in center position suggests that there may be a strong attentional component involved in participants' response times, and that the center position may have been privileged in attracting participants' attention. Apparently, participants looked at and planned to say the center word, only making a saccade to one of the flanking stimuli to the left or right if cued to do so.

This result was not expected, as participants were told to note all three potential target words, not just the center. However, without additional data, such as eye-tracking information, it is impossible

to be certain where participants directed their gaze. Nevertheless, this unexpected effect of stimuli position suggests that in many of the experimental trials, participants may not have noticed that a competitor to the target word was present on screen. If the participants really did simply plan to say the word in the central screen position until cued otherwise, then competitors would only have an effect when in this privileged center position.

With this in mind the collected data was repartitioned into the following five conditions:

1. **Target Off-Center, Competitor Center**
2. **Target Off-Center, No Competitor On-screen**
3. **Target Off-Center, Competitor Off-Center**
4. **Target Center, No Competitor On-screen**
5. **Target Center, Competitor Off-Center**

The means of these new conditions are shown in Figure 15. If the center position was privileged, it is predicted that condition 1 would have the slowest reaction times, since subjects presumably spent more time planning the competitor word. Conditions 2 and 3 would be about identical with somewhat faster reaction times, and conditions 4 and 5 would be the fastest due to the target being in the privileged position. It is also predicted that VOT lengthening should be correlated with the expected reaction time pattern.

Two separate MCMCglmm analyses were performed using these new conditions. In the first analysis, the fixed effect was defined only over the subset of conditions where the target item was in either the left or right position (conditions 1-3 above). Condition 2 (No Competitor Onscreen) was used as a baseline to which conditions 1 and 3 were compared. Results are shown in Table 16. As expected competitors have a significant effect on reaction times and VOT only when in center

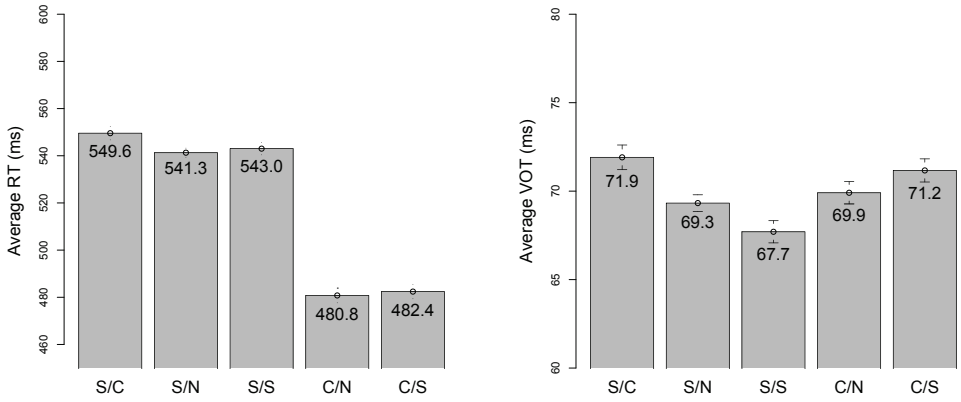

Figure 15: Experiment 6: Latency and VOT broken down by relative placement of target and competitor. S/C: target on the side, competitor in the center, S/N: target on the side, no competitor, S/S: target on the side, competitor on the side, C/N: target in the center, no competitor, C/S: target in the center, competitor on the side.

position. If a competitor is not in the center position, behavior is not significantly different from when there is no competitor onscreen. It is as if participants pay no attention to the competitor at all in that case.

Table 16: Experiment 6: Statistical results after reconditioning.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Target flank/Competitor center (Reaction Time)	10.8966	0.5873	19.8171	< 0.03*
Target flank/Competitor center (VOT)	2.2232	0.07698	4.50029	< 0.04*
Target flank/Competitor flank (Reaction Time)	2.25329	-7.1068	10.4539	< 0.8
Target flank/Competitor flank (VOT)	-0.62601	-2.93183	1.30241	< 0.6

The second model was constructed only over the subset of data where the target *was* in center position (conditions 4 and 5). Treatment coding was used, with the baseline condition again set to the case where no competitor was present. Results from this second model are shown in Table 17. When the target is in center position, there is no significant effect of competitors on either reaction times or VOT. This suggests that participants were indeed paying attention mainly to the center position, and were able to ignore flanking competitors.

Together, these two analyses suggest that in this experiment, the center position was indeed

Table 17: Experiment 6: Statistical results after reconditioning.

Condition	Effect Mean	Lower 95% CI	Upper 95% CI	p
Target center/Competitor flank (Reaction Time)	2.414	-8.283	15.652	< 0.7
Target center/Competitor flank (VOT)	1.6497	-0.8255	4.0361	< 0.2

privileged. Only competitors in this special center position were shown to cause significant slowdown in reaction times, or lengthening of VOT. The two analyses also suggest that reaction times are correlated with VOT lengthening, since both effects pattern similarly in the analyses presented. An important drawback to using two separate analyses for situations where the target is off-center and center is that we end up overlooking the very large reaction time advantage conferred to targets in the center position. One possibility is that the 50ms additional delay required to name targets that are not in the privileged center position may be due to processes external to phonological planning, such as setting up and executing a saccade. If this additional time is independent of the time required to overcome any competition and finalize a phonological plan, it may not be reflected in VOT measurements.

While it seems that the center position is able to draw participant attention better than the side positions, *why* this is the case remains an open question. It also brings up the possibility that the results of Experiments 1-5 were also affected by undetected attentional effects. I would suggest that that attentional effects such as those seen in Experiment 6 depend strongly on the presence of time pressure absent from Experiments 1-5. In particular, by attending to the center position, participants may minimize the average time required to saccade to a target stimulus, thus minimizing their overall reaction times. Since there was no time pressure component in Experiments 1-5, there would be no need for such strategies.

2.3 Summary

Overall, Experiments 1 through 5 indicate that online enhancement effects in speech production are induced when there is competition between a target utterance and a sufficiently similar competing utterance. Effects drop off rapidly as distance increases beyond approximately one phonological feature. Experiments 2 and 5 indicate that competition occurs at a position-specific level of processing, but do not rule out the possibility of whole-word competition. Experiment 6 confirms previous findings indicating that competition affects speech planning latency in a similar way as it appears to affect phonetic enhancement. It also reveals the methodological importance of participant attention in psycholinguistic speech production experiments, and the methodological difficulty involved with examining both planning latency and phonetic enhancement within the same experiment. One of the main goals of the modeling approach presented in Chapters 3 (see Section 3.5, especially) and 4 (see Section 4.2, especially) will be to account for the experimental results presented in this chapter.

3 A Bayesian Model of Speech Production

This chapter presents a Bayesian model of speech production that serves as a complement and potential alternative to previous connectionist models. The model maintains one of the central ideas in speech production theory — the speech production system consists of a number of related levels of processing, each responsible for choosing representations of a particular type — but formalizes the interaction between these levels and the competitive processing within them in a novel way using ideas from information theory and Bayesian mathematics.

The chapter is divided into the following sections. First, a formal description of the Bayesian model is provided, along with a comparison to previously discussed connectionist architectures. In subsequent sections, the model is used to account for a number of the *qualitative* empirical generalization described in the introductory chapter and Chapter 2. The applications discussed include explaining inhibitory and facilitatory contextual chronometric variation (i.e., results from priming and Stroop tasks) as well as explaining the link between planning time and hyperarticulation discovered in Chapter 2.

While a partial analysis of the Experimental results of Chapter 2 is provided in Section 3.5 of this chapter, further development of the model is necessary before a full account can be provided. In particular, the simplifying assumptions used in this chapter limit the model to 'whole-word' levels of processing, while the results in Chapter 2 indicate the importance of competition at 'position-specific' levels of processing. This shortcoming is remedied in Chapter 4, where a position-aware phonological buffer level is added, and a complete account of the experimental results is presented in Section 4.2.

3.1 Model Description

The model presented here follows the the increasingly common application of Bayesian modeling in perception research (Knill & Richards 1996). Bayesian models have been productively applied to aspects of visual perception (Girshick et al. 2011; Simoncelli 2009; Stocker & Simoncelli 2006), written word recognition (Norris 2006; Kinoshita & Norris 2009), and spoken word recognition (Feldman et al. 2009; Norris & McQueen 2008). In perception modeling, the mental system interprets noisy signals gathered by the senses, updating internal beliefs about the state of the external world as more and more evidence accumulates.

The Bayesian word production model developed here, shown schematically in Figure 16, inverts this structure. In Bayesian word production, the signals of interest originate and are processed wholly within the mental system. Instead of interpreting noisy signals from the external world, the levels of processing/representation studied here interpret noisy messages from other levels. Each level maintains a probability distribution over representational states, receives noisy messages from one or more other levels indicating which state it should adopt, and in turn sends noisy messages to other levels.

The message-passing between levels of processing is formally described as a communication system bound by the rules of information theory (Shannon 1948). Noise is an ineluctable feature of any such communication system: noise is present in a signal regardless of whether that signal originates externally (from the environment, or the senses) or internally (from another mental level). One of the simplest, albeit not always the most efficient, approaches to successful transmission over a noisy channel is to use a *repetition code* (MacKay 2003). Repeated sampling in perception can result in a more accurate representation of the external world, as noise in any one sample is averaged out over many samples. For the same reason, repeated transmission of the same message to a level

of pro

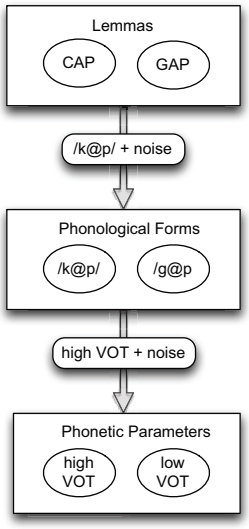

Figure 16: Bayesian Word Production Schematic

Following standard assumptions about the levels of processing that are involved in speech production, it is assumed that there are levels dedicated to selecting concepts, selecting a lexical item (or lemma) associated with a particular concept, selecting a phonological form for the lemma, and finally planning the phonetic/articulatory production of the phonological form. To further clarify the model, the first step will be to explain how the message passing process works, using the link between the lemma and phonology levels as an example. The construction of a message is shown schematically in Figure 17. Each possible lemma can send a characteristic message consisting of a phonological feature vector. The simulations reported here used phonologically realistic feature representations, but for reasons of space we show only part of each vector in the figure. For full assumed feature specifications by phoneme, see Tables 43 and 44 in the In the construction of a message, first one lemma is sampled from the lemma distribution. The characteristic message of that lemma is then corrupted by additive noise and passed to the phonology level.

Upon receipt of a partially corrupted message, a level must have some means of decoding its

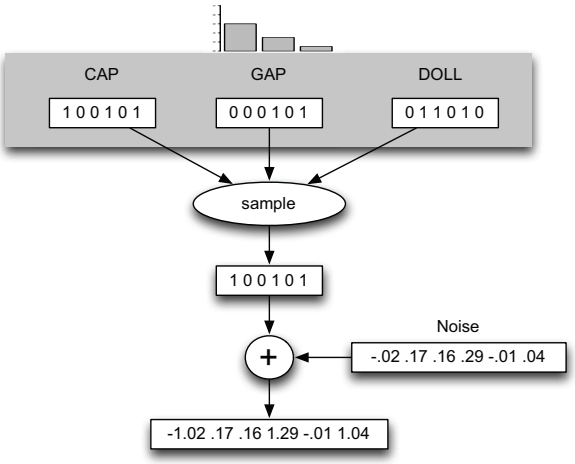

Figure 17: Message Construction: Sending a Message from Lemma to Phonological Levels

contents while minimizing distortion due to noise. Bayesian updating provides a rational mechanism for achieving this goal (Cover & Thomas 2006). The receipt of a noisy message by the phonology level, and the way in which the message is used to update the distribution over word forms at that level, is shown in Figure 18. Each phonological form represented by this level expects a particular message. The difference between this expected message and the message received is passed through a likelihood function to determine the probability that the message received corresponds to that particular form, $p(\text{message}|\text{form})$. Following evidence that the visual perception system can accurately estimate the noise characteristics of visual signals in order to make more accurate inferences, the form of the likelihood function is determined by the type of noise that corrupts the message (Stocker & Simoncelli 2006; Simoncelli 2009). In all simulations reported in this chapter, it is assumed that noise in the word production system has a multidimensional Gaussian distribution, with one dimension assigned to each feature, and all dimensions assumed to be inde-

pendent.[19] Thus, likelihood functions also take on a Gaussian form with similar parameters. Even without knowing the exact levels of noise in the speech production system, we can estimate the parameters of the likelihood function at the phonological level. The experimental results presented in Chapter 2 suggested that significant similarity between forms is defined narrowly — competitors are only sufficiently similar to induce hyperarticulation in target utterances if they differ by about one phonological feature. We can model this behavior by setting the variance parameter of the Gaussian likelihood function so that most of the probability mass is focused tightly around each target representation. The rapid falloff of the Gaussian "bell" shape is conducive to this usage. Using a likelihood value, and the prior probability of each representational form, the level's probability distribution is updated according to Bayes' Rule:

$$p(\text{form}|\text{message}) \propto p(\text{message}|\text{form})p(\text{form})$$

An important feature of the model is that the probability distribution of a level is always renormalized after each Bayesian update. Since the total probability is fixed (all individual probabilities must sum to one), so that taking probability away from one representational state necessarily passes it to other, competition between representational states is implicit in the model. The Bayesian framework employed by the model provides a rational way to integrate evidence in the form of messages over time in order to decide which representational state should ultimately end up with the most probability.

When simulating word production using the model, a phonological form is selected (chosen for production) when it passes a high threshold probability. In the simulations reported here, the

[19] However, other types of noise have been found to produce similar results (e.g., random flipping of binary feature values)

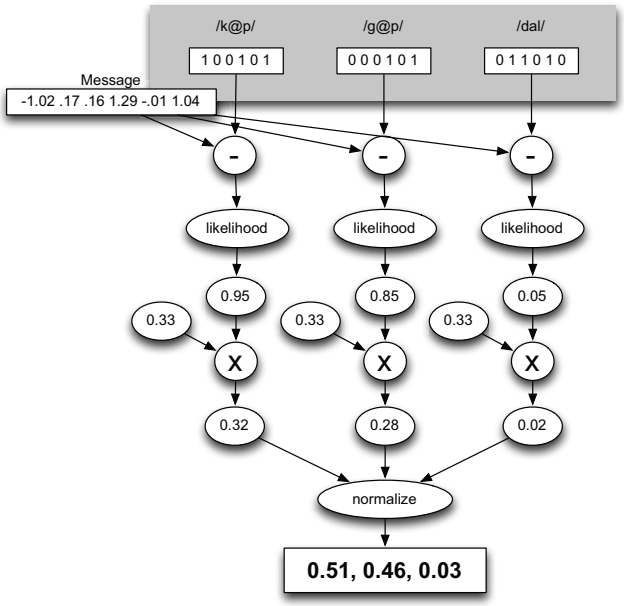

Figure 18: Bayesian Belief Updating at the Phonological Level Upon Receipt of Message from the Lemma Level

threshold is set to 0.95, which means that a form can be chosen only when it has 95% (or more) of the total probability after an instance of the Bayesian belief update. In most situations, a single message will not provide sufficient evidence for any form to reach this threshold after a single update. The necessary level of evidence is accumulated through multiple messages over time. This approach is similar to the Bayesian evidence accumulation using Norris' Shortlist B model of speech perception, which integrates evidence from auditory signals over time. It can also be seen as a multi-class generalization of the evidence accumulation diffusion models used by psychologists to model binary decision making. In these binary models evidence comes in the form of positive and negative numbers. Each piece of evidence is added to a running sum until a positive or negative threshold is reached (Zhang et al. 2012; Teodorescu & Usher 2013; Ratcliff 2013).

A temporal repetition code for communication among mental levels leads to accurate word form selection with high probability, and lends itself well to accounting for latency and other effects observed in production. Since the number of messages required to reach a decision is directly related

to the time it takes the model to decide the appropriate representational state for a level processing, the model uses the Bayesian formalism to describe the speech production system mechanistically, rather than just at Marr's computational level.

In some cases, a module may receive conflicting messages about which representational state it should adopt. This situation is especially plausible whenever multiple upstream levels of processing are sending messages to a common downstream level. During reading out loud, for example, there may be a conflict between messages from a phonological lexicon indicating that a word should be read in a particular irregular way, and messages from grapheme-to-phoneme conversion level indicating what the word's pronunciation would be if it were regular. If this conflict persists, it may not be possible for any particular form to reach the 0.95 probability threshold required for selection, and an alternate response criterion may be needed. In these cases, the most probable form can be selected after the posterior distribution over all possible forms stops changing — that is, if the new distribution is sufficiently similar to the old distribution within a certain tolerance after new evidence is received. This criterion represents a stable state in which probability is split between several forms being advocated for by conflicting messages. Due to the Bayesian nature of the system, choosing the most probable form after this steady state is achieved is guaranteed to be the most rational choice given the (conflicting) evidence received.

3.1.1 Relationship With Connectionist Architectures

Compared to the connectionist models described in the introductory chapter, the Bayesian model described here is most closely related to a cascading architecture with facilitatory links between levels of processing and inhibitory links within levels of processing. The model is cascading since upstream modules may send messages to downstream modules before they have finished deciding upon a representation. Inhibition is affected by the normalization that takes place after every

Bayesian update. Since probabilities must sum to one, normalization ensures that increasing the probability assigned to one representation necessarily decreases the probability assigned to another.

While the model behaves as if it has inhibitory links, there are no actual connections between different representations in a level of processing. In this way, it is similar to Dell's (1986) model of speech production, and shares the benefit of parsimoniously accounting for similarity effects without needing to specify arbitrary connection weights between representations. However, unlike Dell's model, in which similarity effects are caused by feedback from shared components in other levels, the effects of similarity in the Bayesian model are achieved by varying the likelihood function that allows levels to interpret external messages. Each level maintains a distribution over all possible representations at that level. When a message is received, the likelihood of each representation is calculated and probability shifts to those representations that most closely match the message. Similar representations will match similar messages. The Bayesian model instantiated here places a strong emphasis on the role of features in determining the similarity between different representations, a variation of the idea of shared components in Dell's model. Breaking representations down into sets of features allows the creation of tractable likelihood functions based on the number of mismatching features between representations. If similarities could not be broken down in this way, it would be necessary to specify arbitrary likelihood values for every pair of representations. In a connectionist model, this would require every form in one level to be connected to every form in another level, with a weight on each connection hand-picked to correspond to the similarity between the connected pair of representations.

3.2 Inhibition in Chronometric Studies — An Example

As discussed in the introductory chapter, in certain priming tasks, similarity between target utterances and competitors results in *delayed* (longer) response latencies (Meyer & Gordon 1985; Yaniv

et al. 1990; Roelofs 1999). One example is the plan-switching task (Meyer & Gordon 1985), in which participants are prompted to plan to say one form (e.g., the syllable **UP**), but are sometimes cued to say an alternative (e.g., the syllable **UB**). The findings from this task are summarized in Table 18. When the target response is highly similar to the prime response, the time to initiate the target is lengthened. This effect drops off rapidly with increased phonological/phonetic distance. Only alternative responses that are about one feature away from the target seem to induce a significant delay.

Table 18: Plan Switching Task: Similarity = Higher Latency

Planned	Alternative	Difference	Latency
UP	UB	**voicing**	high
UP	UT	**place**	high
UP	UD	**voicing + place**	low

The Bayesian model provides the following qualitative account for this. There are two relevant conditions. In the first case, schematized in Table 19 and Figure 19, the participant must plan to say a prime utterance (the syllable **UP**), but is given a cue to say a different but similar target (the syllable **UB**) instead. Initially, the distribution of forms at the phonology level favors the target utterance, as the speaker has spent some time planning it. After the cue, this level begins to receive messages favoring the target. Since the prime and the target are very similar, the likelihood function favors both of them, and the posterior distribution after each message is received is only slightly different from the prior distribution. Thus, it takes many messages (i.e., higher latency) for the target to reach the threshold probability required for production.

Table 20 and Figure 20 schematize the case when target response (**UD**) is substantially different from the prime(**UP**). Once again, the initial distribution at the phonology level favors the prime. This time, however, the likelihood function responds differently to the messages received after the response cue. Since the prime and target are substantially different, the likelihood favors the

Table 19: Similar Alternative - Plan **UP** with potential alternative **UB**: Each message causes small posterior change.

	UP	**UB**
1) Initial state	0.75	0.25
2) **UB** message likelihoods	0.85	0.95
3) Updated state	0.73	.27

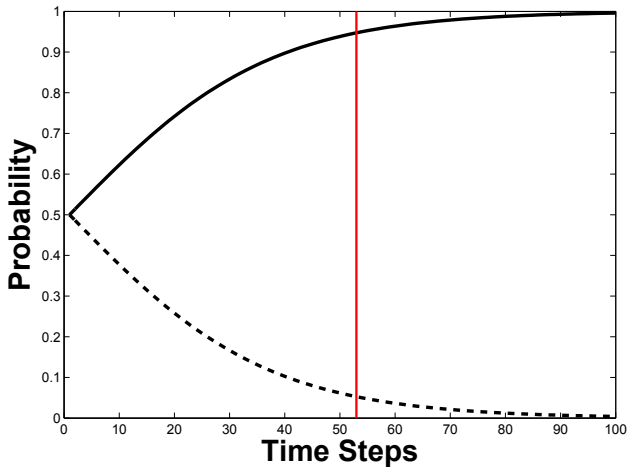

Figure 19: Plan-switching between similar prime and target. Solid line shows the probability trajectory of the form to be produced after cueing. The red line indicates the time step at which a decision can be made. This figure was generated by sending repeated noiseless messages to the phonological form selection level.

target but not the prime. As a result, the posterior distribution after each message is received is more significantly shifted. Since the posterior distribution experiences a larger change with each incoming message, it takes many fewer messages — hence less time — for the target response to reach threshold probability.

Table 20: Non-similar Alternative - Plan **UP** with potential alternative **UD**: Each message causes large posterior change.

	UP	**UD**
1) Initial state	0.75	0.25
2) **UD** message likelihoods	0.25	0.95
3) Updated state	0.44	.56

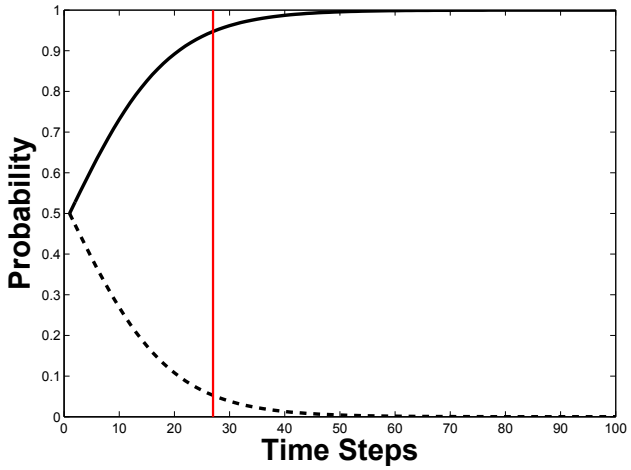

Figure 20: Plan-switching between unrelated target and alternative.

Overall then, latency is higher when the alternative response is more similar to the target, since both the alternative and the target are favored by the likelihood (i.e., there is evidence to produce both forms).

3.3 Facilitation in Chronometric Studies — An Example

Unlike in plan-switching tasks, phonological similarity seems to play a facilitatory role in cue-distractor Stroop tasks (Gordon & Meyer, 1984; Galantucci et al., 2009; Roon, 2012). In a cue-distractor task, participants are taught to associate a visual cue with a particular verbal response (e.g., the syllable **KA** or **GA**). Upon receiving the cue, the participant attempts to produce the associated response as quickly as possible. However, before the participant is able to initiate speech (e.g., at 200ms after the cue), an auditory or visual distractor is presented (e.g., the syllable **PA**).

In spite of the fact that the subject has been given instructions to ignore the distractor, it has an effect on response latency as summarized in Table 21. It seems that when the distractor is sufficiently similar to the target response, production is facilitated relative to the case when the distractor is at a greater distance. However, it is always the case that the presentation of a distractor,

no matter how it is related to the target, results in some production delay relative to the no-distractor case.

Table 21: Cue-Distractor Task: Similarity = Lower Latency

Response	Distractor	Difference	Latency
KA	none	**NA**	minimal
KA	GA	**voicing**	low
KA	TA	**place**	low
KA	DA	**voicing+place**	high

The Bayesian model accounts for this as follows. Again, there are two relevant conditions. Referring to the specific experimental setup used in Roon (2012): depending on a response cue, the participant must say either **KA** or **GA**. For the purposes of this discussion, it is assumed that the **KA** cue is given, and some time has passed so that the distribution at the phonology level has shifted in favor of **KA**. In the first case, schematized in Table 22 and Figure 21, some time after the response cue the participant is presented with a distractor (**PA**) similar to the target, and a few messages corresponding to the distractor are sent to the phonology level. The exact number of distractor messages sent would be determined by the ability of the speaker to control their attention; to ignore the distractor as best they can. Attentional control is beyond the scope of this dissertation. Since the distractor is similar to the target and different from its competitors, the likelihood function provides high evidence for the target and low evidence for any competitors, resulting in a favorable shift in posterior distribution. Note that if the message received corresponded to the target exactly and not just a similar distractor, the target likelihood would be even higher, and the distribution would shift more favorably. Hence, latency is lowest when there is no distractor.

In the second case, schematized in Table 23 and Figure 22, the distractor presented after the cue (**BA**) is substantially different from the target, but similar to the alternative response. The distractor messages now provide low evidence for the target and high evidence for its competitors, causing the posterior distribution to shift in the wrong direction. Correcting this shift requires collecting more

Table 22: Similar distractor - **PA**: Distractor message provides more evidence for target than competitors.

	KA	**GA**
1) Initial state	0.75	0.25
2) **PA** message likelihoods	0.85	0.25
3) Updated state	0.91	0.09

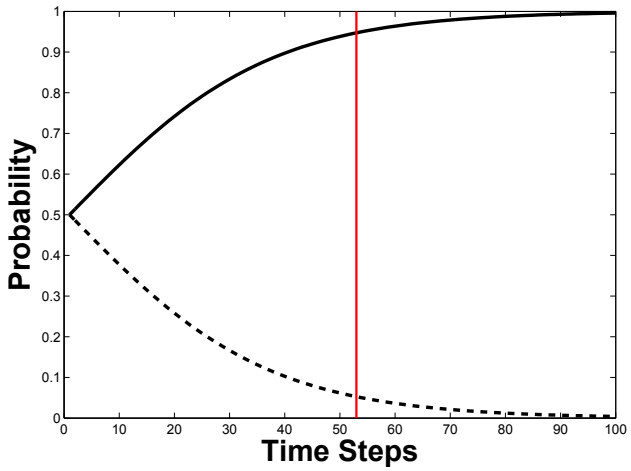

Figure 21: Cue-distractor task with similar target and distractor. Solid line shows the probability trajectory of the form to be produced after cueing.

evidence for the target, resulting in greater latency.

Table 23: Non-similar distractor - **BA**: Distractor message provides more evidence for competitors than target.

	KA	**GA**
1) Initial state	0.75	0.25
2) **BA** message likelihoods	0.25	0.85
3) Updated state	0.47	0.53

In sum, a non-similar distractor causes a larger delay than a similar distractor because it provides strong evidence for the target's competitors and creates a shift in posterior probability towards them which must be overcome.

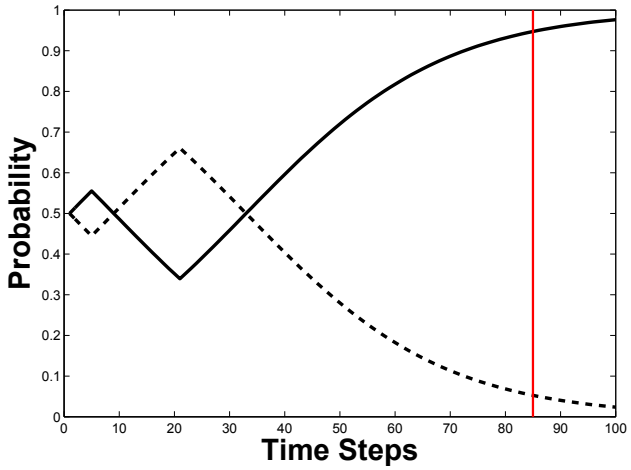

Figure 22: Cue-distractor task with unrelated target and distractor.

3.4 A Unified Framework for Chronometric Facilitation and Inhibition across Tasks

Depending on the task, the particular similarity relationships between the messages, targets, and competitors can result in either facilitatory or inhibitory effects. In general, relative facilitation will occur when the target form either gains more or loses less posterior probability upon the receipt of each message than its competitors. Since similarity is expressed in terms of likelihoods in a Bayesian system, we can predict whether facilitation or inhibition will occur by observing likelihood ratios.

The likelihood associated with a message is the probability of the message given the representation that may have generated it: $p(\text{message}|\text{representation})$. For clarity, we want to express the likelihood as a function of the generating representation rather than the message, whose value is known after it has been received. We can rewrite the above expression as $L(\text{representation}|\text{message})$. The relevant ratio of likelihoods is the likelihood of target representation over the combined (sum) likelihood of all the competing representations:

$$\frac{L(\text{target}|\text{message})}{L(\text{competitors}|\text{message})}$$

In the inhibitory example case above, this ratio favors the target more when the competitor set consists of an unrelated prime rather than a similar prime, so we get relative inhibition when a similar prime is used:

$$\frac{L(\text{target}|\text{target message})}{L(\text{similar prime}|\text{target message})} < \frac{L(\text{target}|\text{target message})}{L(\text{unrelated prime}|\text{target message})}$$

In the facilitatory example case, the ratio favors the target when a distractor similar to the target is presented:

$$\frac{L(\text{target}|\text{sim. distractor message})}{L(\text{competitors}|\text{sim. distractor message})} > \frac{L(\text{target}|\text{unr. distractor message})}{L(\text{competitors}|\text{unr. distractor message})}$$

The results of both tasks, however, are explained within the same framework.

From these likelihood ratios, we can infer that facilitation and inhibition are always relative to each other. Some probability will always shift to both the target and its competitors upon the receipt of a message. Some type of primes or distractors cause relative facilitation when their presence shifts more probability to the target than another type of prime. Indeed, fMRI results indicate both facilitatory and inhibitory brain activation during picture/word naming tasks, suggesting that behavioral observations reflect a balance of both process types (de Zubicaray et al. 2002).

Note that likelihood ratios only indicate the relative amount of probability shifted onto or away from a target representation upon the receipt of a single message. Thus, they represent the rate at which a probability distribution changes. Generally, receiving more messages that shift more probability to the target results in faster selection. However, the actual amount of time a decision takes is also dependent on the starting point of the distribution, the prior. If the prior strongly favors

competing representations rather than the target, it may still take more messages to shift probability onto the target, even if each individual message causes a larger change to the distribution.

Likelihood ratios appear similar to the activation ratios used in WEAVER++ to determine reaction times. In WEAVER++, after activation spreads throughout a level of processing, and one form is selected by passing an activation threshold, the amount of time required to actually produce it is drawn from a distribution parameterized by the activation of the form divided by the total activation in the level, including the activations of all the form's competitors. Thus, the activation ratios in WEAVER++ are a post-hoc mechanism added to the system precisely for the purpose of generating reaction times. In contrast, the likelihood ratios in the Bayesian model arise from one of its core mechanics: message passing. The message passing mechanism, combined with Bayesian updating, allows reaction times to be generated directly.

3.4.1 Understanding Priming Results

As discussed in the introductory chapter, priming studies tend to follow several common patterns. The Bayesian model accounts for these through the manipulation of prior probabilities after each production. This requires the assumption that after an utterance is produced, the prior probability of the representations corresponding to that utterance is increased at all appropriate processing levels for a short period of time. This change in prior affects subsequent productions. The assumption that producing an item once increases the chances that it will be produced again is borne out in statistical examinations of corpus data (Madsen et al. 2005). Thus, it is reasonable for the speech production system to have evolved to change prior probabilities in this way.

The first generalization mentioned with respect to priming was that identical primes have a facilitatory effect on target production. In the Bayesian model, this is attributed to the fact that the prime shifts the probability distribution in favor of the target, leading to a better starting point for

target production than if the distribution was skewed towards another competitor.

The second generalization was that increased similarity (except identity) between the prime and the target leads to slower productions. This is an instance of the inhibitory example case described above. Producing/planning a similar prime raises the prior associated with that prime, and lowers the prior associated with the target. Attempting to produce target requires overcoming this prior skew, but the process is slower than it could be as the likelihood function favors both the prime and the target, and only a small amount of probability can shift from the prime to the target with each message received. Overcoming the prior skew is much faster when the prime is not similar to the target, as the likelihood function no longer favors it.

A common result in the phonological priming literature was that primes that matched the target in onset position resulted in slower productions than primes that matched in the rhyme. This can be accounted for by considering the apparent privileged status of onsets in speech perception and processing. Onsets are more informative of word identity that other parts of the word (Marslen-Wilson & Zwitserlood 1989). Novel different words that share the same onset are more confusable and harder to learn than words that share the same rhyme (Creel et al. 2006). Perea & Lupker (2003) suggest that word onsets are less susceptible to noise in the perceptual system. If onsets are indeed special, then they may also be associated with less noise in production. That is, the noisy channels between levels of processing may transmit information about onsets more clearly than information about other parts of words. If a likelihood function associated with a word form consists of the products of the likelihoods of its individual parts, as is the case with multidimensional Gaussian likelihoods used in this dissertation, then the parameters of the portion of the likelihood associated with the onset would be adjusted to reflect this bias. In effect, words that differ in their onset but shared their rhyme would be less similar than words which shared their onset but differed in their rhyme, since the likelihood function around the onset would be more narrow. As discussed above,

this difference in effective similarity would result in onset-matching primes having an inhibitory effect relative to rhyme-matching primes.

3.4.2 Understanding Stroop Task Results

Like priming tasks, Stroop-like tasks including the picture/word interference paradigm are subject to a few common result patterns. Often, but crucially not always, semantically related distractors inhibit production and phonologically related distractors facilitate production. The phonological generalization is in line with the example facilitatory task described above. Similar distractors shift more probability onto the target than onto competitors, while unrelated distractors do the opposite. The reverse is often true in the semantic case.

The idea of response criteria is also formally incorporated into the Bayesian model through prior manipulation. While priors may change between every trial in an experiment, as they do in priming tasks, they may also reflect global beliefs about the experiment as a whole. For example, if the participant knows that the experiment will only involve the production of nouns, they may drastically lower the prior probability associated with verbs in the appropriate representational levels. This change would persist throughout every trial of the experiment. In fact, determining which subsets of responses are 'excludable' from competition in this way — the ways in which the response set may be cut — may reveal the set of dimensions that define lexical representations

In addition, it seems that lower-frequency distractors have a larger effect on the latency of target productions that higher frequency distractors. This may be attributed to attentional control. As mentioned above, the Bayesian model as currently implemented has no means of determining how long messages associated with a distractor are sent (i.e., how much attention a participant pays them while attempting to plan and produce the target). However, it has been shown that less predictable, more surprising words in a particular context are associated with characteristic brain activity (Lau

et al. 2008) and higher levels of processing (Hale 2001; Levy 2005). It seems possible that they also draw more attention. Thus, lower frequency distractors may interfere with target production more because the brain has more trouble ignoring them.

One remaining question is the extent to which delays during stroop tasks are a result of perception rather than production processes. This question hinges on the extent to which *identifying* the target stimuli is made more difficult by the presence of a distractor.It is difficult to rule out a perceptual component in many tasks, as the Bayesian framework proposed in this dissertation assumes that competition also occurs in perceptual levels of processing. However, the stimuli design in some tasks suggests that a production component must also be involved. In Roon's (2012) task described above, participants were cued to make a particular response using non-linguistic shapes, and then presented with auditory distractors. Assuming that competition is a function of similarity, It seems unlikely that the auditory distractor would interfere with the perception of the cue. Shapes and speech don't have a comparable similarity structure. The relevant competition must occur at level where the non-linguistic cue has already been perceived and converted into a linguistic representation.

3.5 Linking Response Latency and Hyperarticulation

In order to explain the results of Experiments 2-6 in Chapter 2, the model must be able to produce phonetic variation in addition to planning latencies as described above. In particular, it must account for how competition as embodied in Experiments 2-6 induces target-initial VOT hyperarticulation. As discussed in Chapter 2, hyperarticulation was found in competitive conditions where higher response latency was also expected. In addition, both latency (e.g., Meyer, 1985) and hyperarticulation seem to respond to similarity between targets and competitors in a non-linear fashion. In order for competitors to induce increased latency or increased hyperarticulation in target utterances, they

may only differ from them by a small amount (about one phonological feature). As differences increase, effects drop off rapidly. Finally, Experiment 6 tested the general correlation between latency and hyperarticulation directly in a speeded production task, and found evidence of increased VOT in cases of increased latency.

A correlative link between hyperarticulation and latency may be derived from the mechanics of the Bayesian speech production model, making the present model the first in the literature to simultaneously attempt to explain both chronometric variation and phonetic variation at the same time — both arising from the same underlying mechanism.

As shown in Figure 16, the phonology level in the proposed model can be linked to a phonetics level that maintains a distribution over possible phonetic realizations. Following Johnson et al. (1993), these phonetic realizations are assumed to be extreme, or maximally hypo or hyperarticuated articulatory targets (e.g., minimum or maximum VOT targets). Due to the blending effects described below, these extreme pronunciations are unlikely to surface during ordinary speech. Formally, the channel between phonology and phonetics works identically to the channel between lemmas and phonology, or any other pair of connected levels. The phonology level sends messages to the phonetics level indicating which phonetic realization is preferred, and the phonetic level updates its distribution according to Bayes' rule.

The message passing between phonology and phonetics stops when a decision about which form to produce is made at the phonology level (i.e., some form achieves threshold probability). At this point, the phonetic realization of that form can be extracted as a deterministic function of the posterior distribution in the phonetic level. For example, we can consider the case where a voiceless-initial target utterance competes for production with an alternative utterance that differs from it by some number of features. The phonology module must decide between the target and the competitor, while the phonetics module must decide what VOT the final output should have

based on the final distribution over articulatory targets representing minimum and maximum VOT. In experimental data, VOT values for voiceless consonants only vary by small amounts around a high mean value. This suggests that we would like to convert between probabilities and actual VOT values in such a way that even a small preference for high VOT (e.g., p = 0.6), would result in a large amount of VOT (e.g., 90% of the maximum possible value) being expressed. Hyperarticulation caused by a larger preference for high VOT should only make small absolute changes to expressed VOT (i.e., effects should be in the the range of 90% to 100% of maximum possible VOT). One possible transformation function that can convert between probabilities and VOT values while maintaining these properties is as follows:

$$\%\text{ of maximum VOT} = \frac{p(\text{max VOT})^T}{p(\text{max VOT})^T + p(\text{min VOT})^T}$$

where T is a free parameter to be fit, traditionally referred to as temperature due to the relationship between this function and similar functions used in statistical physics, and % maximum VOT is defined relative to the particular phoneme being produced.

Figure 23 shows the results of a series of simulations that varied the distance between the target utterance and its closest competitor in the salient-competitor paradigm of Baese-Berke & Goldrick (2009). As feature distance increases, there is a rapid drop-off in both the time it takes for the phonology level to settle on the target form and the value of the phonetic parameter associated with the form. This pattern arises with a variety of model parameterizations with respect to noise and likelihood functions.

Crucially, decisions at the phonology level take longer when there is a greater amount of competition (e.g., when many competitors are very similar to the target). These longer planning times allow more messages to be sent from the phonology module to the phonetics module, so the phonetics module will ultimately be presented with a greater amount of evidence for the max VOT target.

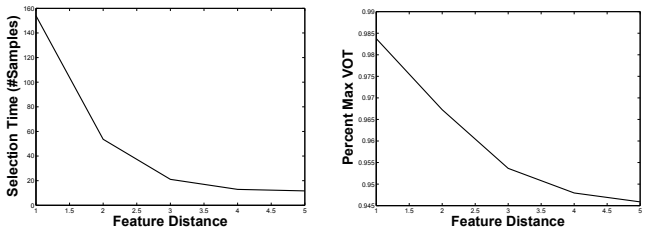

Figure 23: Simulation Results: Selection time and VOT hyperarticulation as distance between target and competitor varies from 1 to 5 features.

This results in a more skewed posterior distribution and ultimately a longer VOT value. Thus, this modeling approach asserts that hyperarticulation is a consequence of longer times spent planning speech, which in turn are characteristic of higher difficulty and competition during the process of speech production. Because the level of hyperarticulation tracks production latency, both end up with a similar behavioral profile. In particular, both types of effects show a similarly rapid drop as feature distance between competitors increases.

This result allows us to describe the competitive behavior discovered in Chapter 2. First, we assume that the presented experiments (all of which shared essentially the same overall design) map onto the Bayesian model as follows. At the start of each trial, the speaker is presented with three words, any of which could be a production target with equal probability. In turn speakers narrow their prior beliefs at the phonological level, assigning a probability of $\frac{1}{3}$ to each possible output form. When the target word is highlighted (and when attention is paid to it), the phonological level begins to receive messages corresponding to it. Shifting a large amount of probability to the target form (starting from the initial uniform prior distribution) takes more or less time depending on how similar to the target the onscreen competitors are. If the competitors are very similar, more messages will be required because the competitors will also be favored by the likelihood function (a scenario similar to the inhibitory priming example discussed above). It is further assumed that after the target becomes known, competitors act as primes rather than distractors. As speakers have

had ample time to see and process the competitor words, they are no longer surprising and are easy for the attentional system to ignore. As shown in this section, increased hyperarticulation is a mechanical consequence of longer planning times in the Bayesian model of speech production. Thus, competition in the experiments of Chapter 2 leads to increased hyperarticulation via increased latency.

Unfortunately, the account presented so far is not complete. While it shows how increased competition can lead to increased hyperarticulation, it does not explain the position-specific nature of competition discovered in Chapter 2. In particular, the results of the Experiments 2-6 in Chapter 2 indicated that word onsets were hyperarticulated when there was significant competition during the phonological planning stages, but only when this competition could be localized to the onset position. For example, VOT lengthened in the case of CAP versus GAP, but not in the case of CAP versus CAB. However, the only competition possible in the Bayesian model as described in this chatper is between whole-forms. Thus, both CAP versus GAP and CAP versus CAB should take about the same time to resolve, since both competitive scenarios feature minimal pairs differing by the same amount (about one phonological feature, modulo the potential special status of onsets). This implies that the level of hyper-articulation in both cases would be identical, a result which is not borne out by the data. Resolving this problem requires the addition of an additional, position-specific level of phonological processing to the model. This necessary extension is presented in Chapter 4.

3.6 Summary and Shortcomings

The Bayesian model presented in this chapter is able to provide a unified account of inhibitory and facilitatory effects in contextual studies of chronometric variation in speech production. It also describes a link between planning latency and hyperarticulation. Examining the model's predictions

with respect to chronometric and phonetic variation induced by lexical factors such as neighborhood density, its predictions with respect to error patterns, or how it can be extended to multi-word utterances is left to future work; some possible approaches are discussed in Chapter 5.

As described so far, the model has one major drawback. It only includes "lexical" levels of processing that maintain distributions over a finite set of holistic representations. This is particularly inadequate for describing the full range of phonological processes, including the position-specific competitive effects observed in Chapter 2. If each level of processing can only sustain competition between whole-word forms, then there is no way for competitive effects to manifest in a particular position of the word. Chapter 4 addresses this shortcoming by including a "post-lexical" processing level to the model. This additional level describes competition at multiple positions in the phonological string under construction.

4 Model Extensions

4.1 Making Room for Novel Utterances and Phonotactics

The Bayesian model of speech production presented in the previous chapter relies on various simplifying assumptions for expository clarity. In particular, the model uses simple discrete probability distributions over finite dictionaries of forms. That is, it only incorporates 'lexical' levels of processing. This limited representation is inadequate for describing the full range of human speech capabilities. As discussed in the following section, it prevents the model from fully accounting for the experimental results of Chapter 2. In addition, people can produce previously unknown non-words, and, a speaker's phonotactic knowledge bears on how easy or difficult it is to plan and produce known and especially novel utterances. Phonotactically ill-formed utterances are (on average) produced more slowly and are more prone to errors (Vitevitch et al. 2004; Vitevitch & Luce 2005). Discrete distributions over finite dictionaries do not permit analysis of these phenomena without becoming intractable (i.e., requiring prohibitively large dictionaries including exponentially many non-lexical forms). This chapter presents extensions to the model that enable it to overcome these limitations. By using a *factored* representation of probability distributions, it becomes possible to compactly describe the probability of any string within a bounded length. Bayesian updates and other inferences involving these new distributions can be performed efficiently by relying on a class of probabilistic graphical models known as factor graphs. In fact, the discussion of hyperarticulation in the previous chapter implicitly assumed that phonetic levels of processing use these factored representations, as individual phonetic primitives were selected in a position-specific manner, rather than complete articulatory plans.

Within the theoretical framework of speech production, this chapter shows how to model the post-lexical level of phonological processing. As described in the introductory chapter, the lexical

part of the phonological encoding process consists of selecting a phonological form for a word from a mental dictionary of phonological forms. This process receives messages from a lemma level of processing, and is identical to the phonological part of the model described in the previous chapter. The job of the newly added post-lexical level is to construct a phonological representation from its component parts. Again following conventions in the speech production literature, I will call this new level of processing the phonological (output) buffer (Harley 2001). The phonological buffer mediates between the phonological dictionary, and processes of phonetic encoding. It may also receive messages from other processes depending on the task, such as orthography-to-phonology conversion, or sound-to-phonology conversion The overall modifications to model structure obtained

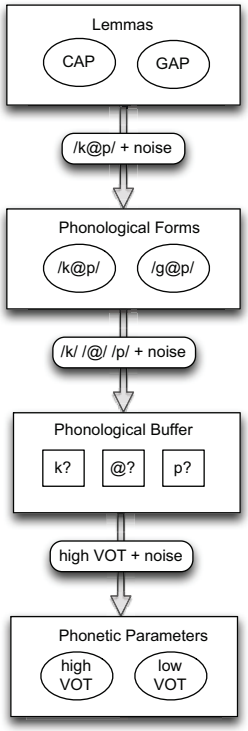

Figure 24: Speech Production with Phonological Buffer

Unlike the lexical phonological level, which can only represent a discrete distribution over a

fixed set of forms, the phonological buffer must be able to compactly represent probabilities over arbitrary strings of phones (within a bounded length, possibly determined by factors like limited working memory). We will denote a string s as $x_1x_2x_3...x_n$, where x_1 is the phone at position 1 in the string, and so on. $P(s)$ is the probability of the string that we wish to describe. We can think of $P(s)$ as a joint distribution over several random variables, X_1 to X_n, one for each position in the string. Each random variable can take on one of N states, where N is the number of phones in the phonological system of the language that can appear in that position. To keep the discussion here simple, and allow for easier visualization, we will cap the phonological buffer at three positions. The construction readily generalizes to longer forms. Thus, we would like to represent the joint distribution over all the positional random variables in the buffer in the buffer:

$$P(s) = P(X_1 = x_1 X_2 = x_2 X_3 = x_3) = ?$$

Normally, specifying a joint distribution over a large number of random variables would require listing one probability for each unique configuration of the variables. If there were n variables, each with N possible states, we would need to specify N^n probabilities, a number which quickly becomes too large for any computer (and likely the brain) to work with. In many cases however, we can take advantage of the fact that some of the variables are independent of each other, or at least independent given the values of some subset of the other variables (known as conditional independence) to represent the distribution in a simpler, factored form. In particular, we would would like to be able to represent the distribution as a product of the following form:

$$P(s) = \frac{1}{Z} \prod_i f_i(\chi_i(s))$$

Where each $f_i()$ is a factor function defined over some subset $\chi_i(s)$ of the random variables in the distribution. Note that in this formulation, there is no need for the individual factors to be normalized in any way. A proper probability distribution is maintained by the presence of the partition function,

or normalizing constant Z, which is defined as the sum over all possible strings:

$$Z = \sum_{s} \prod_{i} f_i(\chi_i(s))$$

The only constraint on the factors is that they be non-negative, as negative probabilities are not interpretable.

Given a distribution in a factored form, we can can represent it with a type of graphical model known as a factor graph. An example graph is shown in Figure 25. A factor graph consists of two types of nodes, variable nodes and factor nodes. Variable nodes represent the random variables in the joint distribution. They are usually depicted as open circles if the value of a variable is unknown, and filled circles if it has been observed. In our case, there will be one variable node per position in the phonological buffer. Factor nodes represent relationships or dependencies between the random variables. They are usually depicted as squares. There is one factor node in the graph for each factor function in the product distribution we are able to define, and it is connected to the subset of variable nodes that the factor function is defined over. Since factor nodes can only be connected to variables nodes, and variable nodes can only be connected to factor nodes, factor graphs are bipartite graphs. Note that the partition function Z is not explicitly represented anywhere on the graph.

Once we have a distribution in factor graph form, we can use a class of dynamic programming algorithms known as belief propagation (so called since the algorithms involve passing messages between the nodes in the graph that themselves resemble probability distributions) to efficiently solve a number of inference problems related to the distribution. For the purposes of this dissertation, the ability to calculate the following is most relevant:

- Given that the values of some of the individual random variables have been observed, what are the marginal probabilities of each of the remaining unknown variables? **Calculated by the SumProduct algorithm.**

- What is the partition function of the conditional distribution of these unknown variables given the observed variables? **Also calculated by the SumProduct algorithm.**
- What is the most probable arrangement of all the unknown variables, given the observed ones? **Calculated by the MaxProduct algorithm.**
- What is the probability of this optimal state? **Can be calculated directly given the factor functions and the partition function.**

If a given graphical model does not have any cycles, meaning that it is not possible to start at a particular node in the graph and find a path that leads back to it, then belief propagation algorithms are guaranteed to converge to the correct answers to the questions above in a time proportional to the graph's diameter (the longest number of edges that must be traversed to get from one node to another in the graph without backtracking or taking detours). If the graphical model does contain cycles, or loops, the algorithms are not guaranteed to find correct answers, or even to converge in finite time. While exact inference can still be performed using slower algorithms such as the graph cuts or junction tree algorithm, it has been discovered that in practice, belief propagation algorithms work remarkably well on general graphs despite the lack of guarantees, and provide a reasonably accurate approximation to the correct inferences. When belief propagation is used in this heuristic way, it is referred to as loopy belief propagation. For a detailed discussion of mechanics of belief propagation and its limitations, see Bishop (2006).

With solutions to the basic inference problems above, we can carry out the functions required by levels of processing in the Bayesian model of speech production. First, a level must maintain a distribution over representational states, and can make decisions after some representational state reaches a threshold probability. Thus, we need to know the most probable state in the distribution at any given time and its probability. Belief propagation gives us the most probable joint configuration

of all the unknown random variables in the distribution s_{max} as well as the partition function of the distribution Z. From this, we can calculate the probability of the most probable configuration:

$$P(s_{max}) = \frac{1}{Z}\prod_i f_i(\chi_i(s_{max}))$$

Belief propagation also allows us to track the marginal probabilities of individual variables in the distribution using the SumProduct algorithm. Thus, we can follow the probabilistic dynamics of individual positions in the phonological buffer.

In addition, each level must be able to send messages to other levels, based on its current distribution. As discussed in Chapter 3, this means choosing a way of sampling from the current distribution. Implementing MAP sampling, or always sending a message corresponding to the most probable current representational state, is trivial using graphical models, since the MaxProduct algorithm lets us calculate the most probable state of the distribution directly. Drawing samples based on their probabilities is somewhat more complicated, but can be implemented using a Gibbs sampling technique. Gibbs sampling is a form of Markov-Chain Monte Carlo (MCMC) sampling that generates samples of a distribution by selecting a single variable in the distribution, fixing all the others to particular values, and sampling from the free variable's marginal distribution given the fixed values of the other variables. Subsequent samples are generated by alternating the variable that remains free while the rest are fixed. Graphical models allow us to find Gibbs samples since we can use the SumProduct algorithm to find the marginal distribution of a single variable while the other variables are fixed. Clamping variables to fixed states requires setting the factor functions in the graphical model to output zero for input variable values other than those we have selected.

Superficially, probabilistic graphical models appear similar to activation-based neural networks. They both consist of a number of nodes connected by edges, and computations in both frameworks involve nodes passing messages to each other along the edges. However, the analogy is limited to these high level observations. For our purposes, it is better to consider graphical models as con-

venient representations of probability distributions that render the relationships between individual random variables explicit, and ease the calculation of various inferences. The message passing used in these inference calculations, as opposed to the message passing between levels of processing described in Chapter 3, is not intended to necessarily have a psychologically relevant interpretation, and can be assumed to take zero time.

In the next section, I complete the analysis of the experiments presented in Chapter 2 that began in Section 3.5 of Chapter 3. This is followed by a proof-of-concept graphical model implementation of the phonological buffer that includes the speaker's phonotactic knowledge. Finally, I present further extensions to the model that leverage graphical model representations to implement active phonological processes including syllabification.

4.2 On the Interaction between Lexical and Post-lexical Phonology

Mediating between the lexical whole-form processing level and the phonetic planning level with a post-lexical phonological buffer results in new competitive behavior that explains the results of the experiments discussed in Chapter 2, but requires some amendments to the basic discussion of model behavior in Chapter 3.

Again, the results of the Experiments 2-6 in Chapter 2 indicated that word onsets were hyper-articulated (via VOT lengthening) when there was significant competition during the phonological planning stages, but only when this competition could be localized to the onset position. If the only phonological competition is at the whole-form level, as it is in Chapter 3, then this position-specific competition cannot be accounted for.

However, when we connect the lexical whole-form processing level discussed in Chapter 3 to the phonological buffer introduced here, we will see that problem can be resolved. In particular, competition at the whole-word level serves as a 'gating' mechanism to competition at the positional

level. Competition at the positional level is only possible given strong competition at the whole-word level, but strong competition at the whole-word level does not guarantee strong competition at the positional level. Ultimately, it is the level of competition at the positional level that governs levels of hyper-articulation and planning latency. Hyper-articulation tracks the marginal probabilities of the segments at each position in the phonological buffer. Once the marginal probability of a particular segment becomes sufficiently high, the phonetic plan associated with that position is fixed. Speech itself can can only proceed once a decision has been made about the contents of every position, or the joint probability of some string reaches a threshold.

Three competitive scenarios will illustrate the range of possible behaviors for the combined lexical/post-lexical system. First, the CAP versus GAP case. Since CAP and GAP are a minimal pair, competition between them at the whole-form level will be strong. This will result in many mixed messages being sent to the positional level about the identity of the first phoneme in the phonological buffer (/k/ or /g/). It will take a long time to resolve this positional competition, leading to increased hyper-articulation at the word onset. In this case, we see that strong competition at the whole-form level permits strong competition in certain positions of the phonological buffer.

In the CAP versus CAB case, competition at the whole-form level will once again be strong since the two competitors are a minimal pair. However, both words send the same message to the initial position in the phonological buffer — that the onset phoneme is /k/. Thus, even though competition at the whole-word level takes a long time to resolve, there is practically no competition at the positional level with respect to the onset phoneme (other that any delays incurred due to phonotactic priors). Little competition in the onset position of the phonological buffer leads to little hyper-articulation, as observed in Chapter 2. In this case, we see that strong competition at the whole-form level does not necessarily lead to strong competition at every position in the phonological buffer.

Finally, in a hypothetical COT versus GAP case, there would be very little competition at the whole-form level since the two competitors do not share many features and it is easy to rule out one of them. Thus, while the onsets of the two competitors differ, the phonological buffer receives very few mixed messages. The whole-form level can quickly decide whether the target is COT or GAP, and thus the positional level quickly begins to receive either only /k/ or only /g/ messages. Thus, from the perspective of the phonological buffer, the COT versus GAP case quickly begins to look like the CAP versus CAB case, and there is little competition in the onset position, leading to minimal hyper-articulation. In this final case, we see that if there is weak competition at the whole-form level, strong competition is not possible at the positional level.

In addition to permitting an explanation of the positional competition effects described in Chapter 2, including both lexical and post-lexical levels of processing in the overall model of speech production allows us to approach other questions that cognitive scientists have focused on in the past. In particular, during the process of reading aloud, the phonological buffer likely receives messages both from the lexical levels of processing where irregular pronunciations are stored, and from grapheme-to-phoneme conversion processes that compute regular readings for written strings. Presumably, both of these inputs are active when reading both known words and novel words or non-words (novel words and non-words serve as evidence for similar known word forms in the lexicon). When reading known words with irregular spellings, we want to ensure that the phonological buffer settles on the correct pronunciation as defined by the lexicon rather than what regular spelling-to-sound rules might imply (e.g., "indict" is pronounced as /ɪndaɪt/ rather than /ɪndɪkt/). This might mean weighing evidence from the lexical and grapheme-to-phoneme levels differently, possibly by assuming a less noisy connection between the lexical level and the phonological buffer and expressing this assumption by manipulating the likelihood function associated with the connection. A full exploration of how this competition between different levels of processing would play

out is beyond the scope of this dissertation, and is left to future research.

4.3 An E

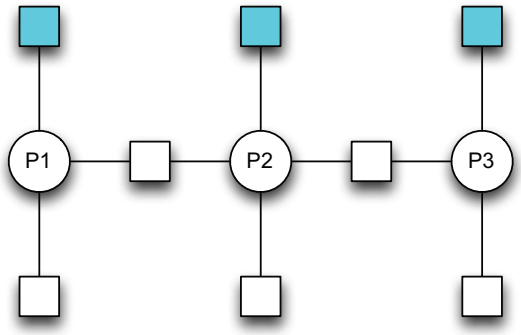

Figure 25: Factor graph model of the phonological buffer. Circles are variable nodes representing positions in the buffer. White squares are factor nodes representing prior phonotactic knowledge. Blue squares are special accumulator factors that represent external evidence about the identity of the phoneme in each position.

Figure 25 depicts a simple graphical model representing the phonological buffer. It includes three variable nodes representing positions in the buffer. Each of these nodes can take on one of 45 states, each corresponding to a possible phoneme of English. The white factor nodes in the graph represent the speaker's prior phonotactic knowledge. The unitary white nodes (those connected only to a singe variable node) represent the prior probability of each phoneme in each position. For the purposes of this demonstration, all unitary factors are set to 1. Effectively, this means that any phoneme is equally likely to appear in any position, assuming the other positions did not exist. The binary white factors (those connecting positions one and two, and two and three) represent prior phonotactic knowledge about which phonemes are more likely to appear adjacent to each other.

In this case, the values of these binary factors were derived automatically from the phonotactic statistics of the words in the Hoosier Mental Lexicon (http://neighborhoodsearch.wustl.edu/Home.asp). First, all of the CVC-structured words were extracted from the lexicon. For each possible bigram

of phonemes in the first and second word position (e.g., /tɑ/, /bɑ/, etc.), the number of words that contained that phoneme combination, or the type frequency of the bigram, was recorded. For each possible bigram, the binary factor in the graphical model was set to its recorded type frequency plus 1. The addition of the extra term was included for smoothing purposes — to ensure that no bigram had absolutely zero probability of occurring. To create the second binary factor, the procedure was repeated, with bigram counts tallied over the second and third word positions in the CVC subset of the Hoosier Mental Lexicon. The frequency of use of each of the CVC words (token frequency) was not considered when calculating the binary factors. This is because type, as opposed to token, frequencies, have been shown to be better predictors of participants' phonotactic judgements (Hayes & Wilson 2008).

Since only CVC words were used to derive the phonotactic binary factors for the example graphical model, we have effectively built a phonotactic grammar that disprefers non-CVC structured inputs. Consonant (CC) and vowel (VV) clusters would receive low probability as they did not appear in the input, and only get non-zero probability thanks to the add-one smoothing used when building the phonotactic factors.

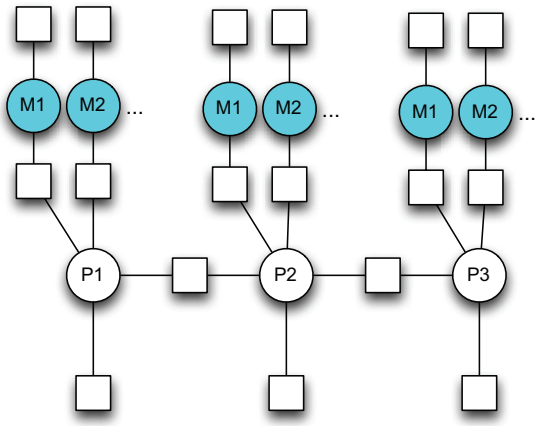

Figure 26: Expanded factor graph model of the phonological buffer. The blue circles represent the messages received by the phonological buffer over simulated time.

Unlike the standard white factor nodes in the factor graph, the blue factor nodes connected to each position variable have a special interpretation. They function as *evidence accumulators*, accumulating any information that the phonological buffer receives about the identity of each phoneme in each position from any external levels of processing. One way to understand how these accumulators work is to think of them as permitting a shorthand version the factor graph shown in Figure 27. This fleshed out graph includes a number of new variable nodes. These new variables represent the messages received by the phonological buffer for each position over simulated time. A new node is added to the graph every time a new message is received. Since the messages are observed, their corresponding variable nodes are filled in, and can only take on one state (the observed message). Two new factor nodes are associated with each observed message node. The unitary factors are clamped to one at the value of the observed message, and zero otherwise. The binary factors connecting the message nodes to the position nodes represent the likelihood of the message given each possible phoneme. As in Chapter 3, messages are instantiated as noisy feature matrices. Likelihoods are calculated by computing the probability of the message according to a multi-variate Gaussian distribution centered on the feature values of each phoneme. In all the simulations presented here, the same phoneme and feature sets were used as in Chapter 3. Through these factors, each of the observed message nodes can send information to the positional variables about which phoneme state they should take on. According to the SumProduct algorithm, the combined information across all the received messages is equivalent to the point-wise product of their phoneme likelihoods:

$$\prod_t \{P(M_t|Ph_i\}_i$$

where $\{P(M_t|Ph_i\}_i$ is a vector representing the likelihood of phoneme the message received at time t for each phoneme i.

Since the likelihoods involved come from multi-dimensional Gaussian distributions, their values

are typically very small. Multiplying many of them together as multiple messages are received can quickly lead to numerical underflow when performing computer simulations using the factor graph. That is, the numbers involved quickly become so small that the computer, with its finite precision, cannot distinguish them from zero. Conveniently, the SumProduct algorithm allows us to normalize each message's likelihoods before multiplying them together, without changing the results of any subsequent inferences performed using the factor graph[20].

$$\prod_t \frac{\{P(M_t|Ph_i\}_i}{\sum_i P(M_t|Ph_i)}$$

This strategy significantly increases the time (number of messages received) required before numerical underflow occurs.

Thus, the combined effect of all the observed message nodes is simply a vector of values, one for each possible phoneme state of each positional variable. This is precisely the same format that a unitary factor defined over each positional variable would have, allowing us to represent the whole set of observed message nodes shown in Figure 27, with the single blue factors shown in Figure 25. Every time a new message is received, we simply update the blue factor by point-wise multiplying the vector of values that it represents by the normalized likelihoods calculated by comparing the newly received message with each possible phoneme. In this way the blue factors act as evidence accumulators: each one compactly represents all of the messages that have been received so far about the corresponding position.

Given the graphical model formulation of the phonological buffer described above, we can now run simulations that show how phonotactic knowledge may affect the time required to plan speech.

[20]In the language of belief propagation, each message node itself sends a message to its associated position node. This message is equivalent to a vector of likelihoods as described in the main text. When performing inference using belief propagation, it is always possible and frequently desirable to normalize messages before sending them, as this helps maintain numerical stability.

Of particular interest are pairs of strings where the same segments are involved, but their arrangement in one case is phonotactically preferred. Since we know from the way that we constructed our phonotactic factors that consonant clusters are dispreferred, one such case is DAD (preferred) versus ADD (dispreferred). We simulated the production planning of both of these words by continually sending noiseless messages corresponding to the phonemes of each word to the appropriate positions in the phonological buffer. All likelihood functions were multi-variate Gaussians with diagonal variance set to 2.

DAD was produced (the string /dæd/ reached threshold probability) in 36 time steps, while ADD took 41 time steps. As expected, poor phonotactics can delay the phonological planning of words that are otherwise equivalent to their phonotactically good counterparts.

4.4 Graphical Models and Phonotactic Grammars

For expository clarity, the discussion above used a simple set of bigram phonotactics derived from the contents of the Hoosier Mental Lexicon. However, the formalism of graphical models permits the plug-in use of much more complex phonotactic grammars (either hand-crafted or automatically learned from the input language). In particular, the Hayes-Wilson phonotactic learner induces a constraint-based harmonic maximum entropy (MaxEnt) phonotactic grammar that can be readily converted into a set of factors (Hayes & Wilson 2008)[21]. The grammars learned by the phonotactic learner have proven successful at describing speakers' phonotactic knowledge. For example, the system accurately predicts speakers' judgements of the phonotactic goodness of English onset consonant clusters, and the relative goodness of vowel harmony patterns in Shona.

[21] Inference on factor-graphs has been labeled a form of constraint-satisfaction problem. Smolensky (1986) describes activation-based neural networks designed to maximize harmony that are structurally similar to factor graphs.

Constraints in a learned phonotactic grammar may be defined over two or three adjacent[22] phone positions in a string. Each constraint looks for a certain configuration of features in each phone it is defined over. If the set of phones have the appropriate feature configurations, the constraint is triggered. The harmony, or goodness, of a particular string is defined as the following sum:

$$h(x_1 x_2 ... x_n) = \sum_i w_i C_i$$

where w_i is a weight associated with constraint number i, and C_i is an indicator function with value 1 if constraint i was triggered and value 0 otherwise.

Hayes & Wilson note that these harmony values may be converted to MaxEnt scores, which are proportional to probabilities, simply by exponentiating them.[23] This transformation is crucial for allowing the system to learn both the constraints and how they should be weighed. So, the MaxEnt score of a string is defined as:

$$e^{h(x_1 x_2 ... x_n)} = e^{\sum_i w_i C_i}$$

From this formulation, we can directly extract the factors associated with a probability distribution. Each factor is just the exponentiated value of a constraint's weighted indicator function, and the joint distribution over a string (modulo a normalizing constant Z) is defined as the product over all factors:

$$e^{\sum_i w_i C(i)} = e^{w_1 C_1} e^{w_2 C_2} ... e^{w_N C_N} = f_1 f_2 ... f_N$$

[22] Adjacency is enforced in order to make phonotactic learning tractable. This limitation may be mitigated by splitting a string into multiple tiers, such as separate vowel and consonant tiers, and to enforce adjacency within tiers. For example, two vowel segments may be adjacent on the vowel tier, but may be separated by a number of consonants in the surface string. The tier approach allows the phonotactic grammar to learn about apparently long-distance patterns such as vowel harmony.

[23] MaxEnt scores only differ from probabilities in that they are not normalized. In order to normalize them, we would need to divide the MaxEnt score of each string by a partition function Z, the sum of all MaxEnt scores.

To reduce redundancy, if multiple constraints are defined over the same subset of phones, we can collapse them all into one factor corresponding to the exponentiated sum of their weighted indicator functions.

Having broken down the joint probability distribution over strings into individual factors, we can build a graphical model. The variable nodes would represent the individual segments in a string, and the factor nodes would correspond to the factors defined above, each with connections to the phone positions it is defined over. In the following section, we will see how this convenient conversion between harmonic grammars and probabilities can be used to instantiate processes such as syllabification and allophonic variation with graphical models. At present, this is primarily meant as an extension of the computational capabilities of the Bayesian model. It is not meant to bear on any specific set of experimental results.

4.5 Syllabification and Other Active Processes

The discussion so far has focused on how phonotactic restrictions affect the speed with which the phonological buffer settles on a particular phone in each string position, given evidence from external levels of processing. All the mechanisms and phonotactic restrictions described can be said to apply to the 'underlying form' of an utterance. The active application of phonological processes that induce allophonic variation in underlying forms (e.g., segment epenthesis, deletion, devoicing, and assimilation) and syllabify and add prosody to the resulting strings, have yet to be discussed. Not much is known about the chronometric properties of allophonic variation, such as whether some phonological rules apply more slowly than others. This sparsity of knowledge is partly due to the fact that it is difficult to design experiments that single out phonological rule application.

The chronometric properties of syllabification and prosodification have have been previously studied, but only to a limited extent. In most production models (but notably not WEAVER++,

which uses an active syllabification process external to its core neural network mechanisms (Roelofs 1997)), after a word is selected for production, a predetermined syllabic and prosodic frame associated with the word is extracted form memory. This frame is then filled with the appropriate segments during a phonological selection and encoding process. Thus, a prosodic structure is not built on top of a string of selected consonants. Rather, consonants are selected so that they fit a pre-determined prosodic frame. Debate about the temporal properties of this process has focused on whether one syllable is filled at a time, or if a larger frame accounting for the entire utterance is filled in parallel. The relevant experimental evidence has been mixed. If syllables are filled sequentially, we might expect longer words with more syllables to take longer to plan. Such differences due to length have not been observed (Levelt 1999) (but, see Meyer (2003) for evidence that they may exist in some cases). As discussed in the introductory chapter, results from implicit priming and the rapid production of similar syllables has been used to argue for sequential syllabification. However, the discussion in Chapter 3 suggests that these effects can also be accounted for by allowing the initial parts of utterances to be privileged.

I suggest that graphical models provide a convenient way of implementing parallel active rule-governed syllabification, prosodification, and potentially allophonic processes, as long as we are not concerned with their precise time course. In particular, they allow us to find the most probable surface realization of the phonological buffer at any point during the segmental selection processes described above. The process of choosing which underlying segments belong in the phonological buffer may be seen as progressively clarifying the input that is fed into active phonological processes. These processes in turn can be thought of as occurring in parallel to segmental selection. In the preliminary formulation described here, they do not have an independent time-course. As soon as the underlying segments in the phonological buffer are known, so are their most likely surface forms.

As an example, I will construct a graphical model that performs syllabification of underlying Berber strings. Berber was chosen for this demonstration because its syllabification has already been described by a harmonic grammar and can thus be easily converted into graphical model form. Berber is famous for words consisting of particularly long strings of obstruent consonants (e.g., *txznt*). However, Berber maintains a rhythmic syllable structure. Some consonants function as syllable nuclei, which the rest function as non-nuclei (the distinction between onsets and codas is not important for the phonology of Berber). Dell and Elmedlaoui (1985) described a simple algorithm that can syllabify any string of Berber (that is, determine whether each segment is a nucleus or not). The algorithm relies on the assumption that each Berber segment has one of eight sonority levels, one being the least sonorous and eight being the most. Furthermore, all Berber strings are assumed to be constrained so that no two segments of the same sonority level may appear in a contiguous sequence; any adjacent segments must have different sonority values[24]. The more sonorous a segment, the more likely it is to function as a nucleus. However, there can be no adjacent nuclei. The algorithm is as follows:

Until there are no more segments to syllabify, repeat the following:

1. Find the most sonorous unsyllabified segment, and designate it as a nucleus.
2. Designate the segments immediately preceding and following the new nucleus as non-nuclei.

This behavior of this algorithm was subsequently captured by Smolensky et. al (2006) by a harmonic grammar including both positively and negatively weighted constraints. The harmony of

[24]In reality, there are Berber words with sonority plateaus. However, these were not treated by the Harmonic Grammar analysis discussed below.

a syllabification was calculated as follows:

1. Each segment designated as a non-nucleus contributed 0 to the total harmony.

2. Each segment designated a nucleus contributed $2^s - 1$ to the total harmony, where s was the sonority of the segment between 1 and 8.

3. Any two adjacent nuclei contributed -2^8 to the total harmony.

Constraint 2 ensured that more sonorous segments were designated nuclei when possible. Constraint 3 ensured that the most harmonious syllabification never included two adjacent nuclei. The most either

the -2^8 pen

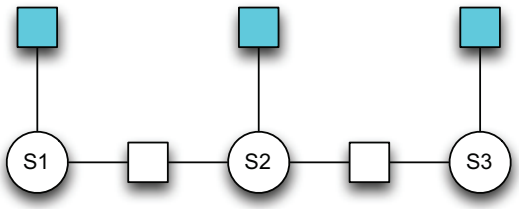

Figure 27: Graphical model used for Berber syllabification. Each random variable corresponds to a position in the phonological buffer and can take on one of two states: nucleus and non-nucleus. Blue factors are a function of the sonorities of the most likely segments in the phonological buffer.

The graphical model that expresses this harmonic grammar is shown in Figure 27. It includes one random variable node for each position in the phonological buffer. Each of these nodes represents the syllabic role of the segment in that position, and can take on one of two states, nucleus or non-nucleus. The model includes unitary and binary factor functions, which are designed based on the constraints in the harmonic grammar. In the harmonic grammar, the correct syllabification has the largest harmony. In the graphical model, this must correspond to the largest product over factors.

We can achieve this relationship by using factors whose values are just exponentiated harmonies.[25] For example, the values of the blue factor corresponding to position one break down as follows:

- e^0 for non-nucleus state
- e^{2^s-1} for nucleus state, where s is the sonority of the most likely segment in position one.

Similarly, the all binary factors break down as:

- e^0 for when the two adjacent positions are not both nuclei
- e^{-2^8} when the two adjacent positions are both nuclei

At any point during the selection process going on in the phonological buffer, we can find the sonorities of the most likely segments in each position, and use them to set the blue factors in the syllabifying graphical model. Using the MaxProduct algorithm, we can then find the most likely configuration of all the syllable position nodes (i.e., whether each position should be a nucleus or not). Because we designed the model to correspond exactly to the harmonic grammar defined in (Smolensky & Legendre 2006), the most likely configuration is guaranteed to be correct as long as the grammar makes the correct predictions. If the most likely segments in the phonological buffer should change, then the most likely syllabification could also change. The final syllabification will thus depend on the final state of the phonological buffer (syllabification itself takes no time in this framework, even though segment selection does).

The example of Berber syllabification is intended to demonstrate that conversion between harmonic grammar and graphical model is a straight-forward procedure. To the extent that processes

[25] Since $ln(e^x e^y e^z) = ln(e^{x+y+z}) = x+y+z$. Taking the logarithm preserves monotonic relationships, so the configuration of x, y, z that maximizes $x+y+z$ also maximizes $e^x e^y e^z$.

other than syllabification, such as allophonic variation, can be described by harmonic grammars, they can also be represented by graphical models. For example, we could design a model with random variables representing possible surface segments, Factors would correspond to the current state of the phonological buffer where the underlying string is being constructed, and to constraints in the grammar controlling when allophonic changes are preferable. The model would then find the most likely surface string given the currently most likely underlying string. Such a model, for example, could be used to explain phenomena such as "wug" testing, in which speakers know that the plural of the previously unseen word "wug" should be pronounced /wugz/, despite the fact that the phonological buffer could receive messages to produce /s/ in the plural morpheme position. Higher level morphological and syntactic processes could also be implemented using graphical models, provided they could be described by harmonic grammars, but this extension is left to future research.

5 Summary, Conclusions, and Future Directions

5.1 Summary of Contributions

This dissertation makes several contributions to the empirical and formal understanding of speech production. On the empirical front, the experimental studies presented in Chapter 2 suggest that online competition among similar word forms during speech production is highly dependent on the nature of their similarity. For competition to have a significant effect on speech production, competitors may only differ minimally (by approximately one phonological feature). Only substantially similar word forms compete, and competition falls off rapidly at greater distances. Furthermore, competitive effects that lead to hyperarticulation seem to occur at a position-specific level of processing. For some particular feature to be hyperarticulated, competition must exist at the position of the that feature (e.g., the initial VOT of CAP may be hyperarticulated if GAP is a salient competitor, but not CAB). As discussed in Chapter 4, this position-specific competition is licensed by whole-word competition, making competition at both lexical and post-lexical levels crucial to an account of the observed effects. Finally, the results of Experiment 6 suggest that there is a correlation between hyperarticulation and latency. In cases where competition is likely to slow down the speech planning process, it is likely to cause hyperarticulation.

On the formal front, a Bayesian approach to modeling speech production was presented in Chapters 3 and 4. The approach does not call for a radical change to the basic processing architecture involved in speech production, as more or less agreed upon by the majority of linguistics and psycholinguists. Relying on the mathematics of Bayesian statistics and information theory, it formalizes how these processes interact with each other, and how competitive dynamics between different representational forms play out within each process. Unlike most Bayesian models in other fields such as speech perception (although see Norris, 2008) which describe data distributions at a

high level (a computational level of analysis), the present model describes the actual mechanisms that lead speakers to behave a certain way, and is thus more appropriately situated at the algorithmic level of analysis.

This probabilistic approach is novel, as the vast majority of speech production models rely on more general neural-network activation-spreading mechanics instead. The adherence of the model to probability theory actually limits the potential range of mechanics and phenomena that can be implemented by it (e.g., the value associated with any representation in a level of processing must be equivalent to a probability, and all probabilities must sum to one, unlike arbitrary activation values). In fact, since a probabilistic model can always be implemented by an activation-based network, the present Bayesian model can be seen as a restriction to a particular subset of activation-based models that conform to the rules of probability.

The Bayesian model of speech production simplifies the analysis of chronometric empirical data, resolving various apparent paradoxes. In particular, the same competitive dynamics explain why competitors sometimes seem to facilitate (speed up) speech production, while at other times they have an inhibitory effect (speech production becomes slower). The correlation between latency and phonetic variation observed in Chapter 2 also falls out of the mathematics of the model. Overall, the Bayesian approach to speech production covers more phenomena than any individual currently implemented activation-based model, yet uses a single uniform process. Furthermore, allowing the model to represent probability distributions as graphical models enables the integration of speakers' phonotactic and grammatical knowledge into the speech production process.

5.2 Current Shortcomings and Future Directions

The work presented in this dissertation suggests numerous potentially fruitful avenues for future research in both the empirical and formal domains.

This dissertation has focused only on modeling and explaining empirical results relating to the chronometric properties of speech, and phonetic variability. The remaining tentpole of speech production research, understanding error patterns, has been left undeveloped. If the modeling approach presented in this dissertation is characteristic of the architecture of the brain, it should also provide insight into when and why people make errors during speech production. In current simulations, any errors made by the model (i.e., times when the wrong form reaches the threshold probability for making a decision before the target form) are rare and simply ignored. Keeping track of how often such situations occur, particularly in response to varied similarity between competing representations, varied bias in likelihood functions or prior probabilities, varied decision thresholds depending on task demands, or varied levels of noise between modules, should yield empirically realistic distribution of errors. Increasing inter-module noise to very high levels would effectively be simulating a kind of 'damage' to the speech production system, and should produce behavior consistent with various cognitive disorders. Increased noise is predicted to result in significantly slower planning times and overall higher error rates. However, whether or not changes to the relative rates of individual error types as a result of the manipulations above match the attested behavior of normal and aphasic speakers is still to be determined.

The focus of the dissertation has also been limited primary to modeling contextual effects (those caused by salient competitors in the speech environment or task). However, as reviewed in the introductory chapter, there is a long tradition of literature dealing with how the relatively static, lexical properties of words, such as frequency and neighborhood density, affect production. As suggested in the introductory chapter, it is possible that these lexical effects have the same source as contextual effects. Because they are usually studied via the production of words in isolation, they may represent the speaker's default prior knowledge in the absence of any contextual information. This idea is supported by the fact that lexical properties only seem to maintain their effect in running

speech when they are correlated with their contextual analogues (e.g., usage frequency only produces reliable effects in running speech when it is correlated with contextual predictability). If lexical effects are indeed just a reflection of default knowledge, then they should readily be explained by the Bayesian model of speech production parameterized with an appropriate set of priors that reflect this knowledge. Simulations of isolated word production would be required to confirm this hypothesis.

While lexical effects might be simply explained the Bayesian model with just an appropriate choice of priors, there are some aspects of the speech production process that would require extensive additions to the model before they could be explained. First, the model is currently limited to the production of short, mono-syllabic and mono-morphemic utterances (in different contexts). Actual speech production involves the planning of much longer phrases (Jaeger et al. 2012a,b; Hilliard et al. 2011). In its present form, the model relies on the simplifying assumption that these long phrases can be broken down into numerous small subparts, each of which can be modeled in parallel and relatively independently. Indeed, this simplification is reflected in much of the empirical investigation of speech production in the psycholinguistic literature, as the vast majority of studies focus on single-word utterances. However, this simplification is obviously inadequate for gaining a complete understanding of speech production. For example, corpus studies of running speech have shown long-range transpositions of segments (MacKay 1970), and chronometric long-range competition effects (slowdown in reaction times when utterances contain many similar segments, even if they are far apart) indicating that while multiple parts of an utterance may be planned in parallel, these processes are not entirely independent (Schnur 2011). It is also the case that the correct surface realization of a linguistic element cannot always be determined without knowledge of distant words in the sentence. The graphical model representational approach discussed in Chapter 4 provides a foothold to understanding how long-distance dependencies might be expressed (as factors

that join distant positions in the graphical models), but the bulk of the necessary development is left as future work.

Attention also seems to play an important role in speech production. Globally, performing a non-linguistic task that requires central attention slows down speech production (Roelofs & Piai 2011). Stroop tasks indicate that speakers cannot help but give some attention to distractors, even if ignoring them would decrease competition and make speech planning faster. In addition, the results of Experiment 6 in Chapter 2 suggest that the salience of a competitor in the speech environment is strongly dependent upon how much attention speakers end up giving to it. Thus, any empirical studies of speech production should take into account how speakers may shift their attention during an experiment. However, the mechanisms that assign and maintain attention are not currently part of the Bayesian model of speech production. The results of any attentional processes are merely stipulated. This is a shortcoming shared by the Bayesian model presented in this dissertation and all current models of speech production.

Finally, the Bayesian model in its current incarnation is designed to explain the chronometric and phonetic variation in speech production, and thus focuses on levels of processing appropriate for the selection of semantic, phonological, and phonetic representations. However, by developing networks of different appropriate processing stages, we can in principle use this modeling approach to describe *any* communicative process in the mind/brain. For example, it might be used to explain the selection of motor programs during writing, or playing music. Since the model can be treated as a general approach to communication, competition, and decision making in the mind, it need not be limited to production tasks. The same basic mechanisms apply in perception, albeit with different modules involved. The field of potential applications is broad enough to be relevant to all corners of cognitive science.

A Experiment Stimuli

This appendix contains stimulus tables for Experiments 1-6 presented in Chapter 2. Aside from the Experiment 1 stimulus table, which uses a special format described below, all tables use a common format. The first column is always the target word that is spoken on a given trial of the experiment. Subsequent columns contain competitors that differ from the target word along some dimension, and two unrelated filler words. Depending on the experimental condition, on a given trial each target word will appear on-screen with either one of its competitors and a single filler, or with both fillers.

In Experiment 1, only vowel properties were measured, and the stimuli are presented as a single table. In Experiments 2-6, the dependent measure was the voice-onset-time of the initial consonant of the target word. Stimuli are broken down into three tables according to the identity of this initial target consonant (/p/, /t/, or /k/).

Words that appeared as targets Baese-Berke & Goldrick (2009) are marked with an asterisk throughout the appendix.

A.1 Experiment 1 Stimuli

Stimuli are formatted as follows. The first column contains words with a lexical neighbor that differs by its vowel. The second column contains paired words that have no such lexical neighbor. The words from column one are presented with either the competitor in column three, or the unrelated filler word from column four. The words from column two are presented with the unrelated filler words from column five.

Table 24: E1: All stimuli.

With Neighbor	No Neighbor	With Neighbor Competitor	With Neighbor Filler	No Neighbor Filler
balk	gawk	buck	rush	cheese
bob	mob	babe	tong	hole
calm	*palm	*comb	dare	fill
catch	match	coach	guess	bead
chip	zip	chop	yoke	work
dash	cache	dish	sauce	psalm
dub	*pub	dab	pace	dual
heave	thieve	halve	rod	house
mash	sash	mesh	gear	pout
paid	jade	pod	lac	hub
pease	tease	pause	gut	phone
pick	thick	peck	lain	soothe
*posh	josh	push	rogue	hag
sought	thought	suit	hope	hair
*tag	nag	tug	heap	take
teeth	heath	*tooth	jack	bole

A.2 Experiment 2 Stimuli

Table 25: E2: /p/-initial target stimuli.

Target	Onset Neighbor	Vowel Neighbor	Coda Neighbor	FillerA	Filler B
*pall	ball	pull	par	nub	thin
pike	bike	poke	*pipe	dart	fog
*pig	big	peg	pit	daft	heart
*punk	bunk	pink	*punt	lard	ship
*punch	bunch	*pinch	pump	wool	*cool
*peat	beat	pet	peak	thug	wren
*peek	beak	peck	*peat	mace	yam
*poll	bowl	pool	*poach	yelp	jam
*palm	balm	Pam	*pomp	nose	hag
pack	back	puck	past	guild	wrist
*pun	bun	pan	*pup	*tar	golf
*pare	bare	pyre	*pale	jug	vet
*pill	bill	*poll	pick	hair	food
*pore	bore	pear	pole	shelf	mile
*putt	butt	pit	puss	zone	land
*pad	bad	pod	pat	norm	van

Table 26: E2: /t/-initial target stimuli.

Target	Onset Neighbor	Vowel Neighbor	Coda Neighbor	FillerA	Filler B
*teem	deem	*tame	teen	mast	board
*tile	dial	tale	time	yap	gong
tent	dent	*taint	*tenth	barn	worm
*tab	dab	tub	*tag	wedge	shoal
*torque	dork	Turk	*torn	wet	loud
*tart	dart	tort	tarp	mouse	gel
*tan	Dan	tin	tad	sheep	hum
*tomb	doom	team	tune	male	whale
*tick	dick	tack	*tiff	size	goon
ton	done	*tan	tug	hill	mouth
tip	dip	*tap	tic	hood	moon
*tame	dame	tome	tape	bench	sign
*tuck	duck	tock	tusk	wit	beard
tote	dote	tight	tone	rug	shaft
*taunt	daunt	tint	*taut	guile	mood
*Ted	dead	tad	ten	cart	goal
*tore	door	tire	toll	goose	half
*teal	deal	tool	*teem	goof	yap
*tyke	dike	take	thyme	sun	bed

Table 27: E2: /k/-initial target stimuli.

Target	Onset Neighbor	Vowel Neighbor	Coda Neighbor	FillerA	Filler B
*cod	God	*cud	*cop	tool	vent
*cuss	Guss	kiss	cull	mar	vain
cap	gap	cup	cat	wolf	dim
*kilt	guilt	cult	*kiln	*toast	yacht
*cuff	guff	calf	*cub	reed	noon
*cab	gab	*cob	can	surf	rice
*kit	git	cut	*kin	lace	bean
cot	got	coat	con	haze	wig
*cape	gape	*coop	*cake	yard	thieve
*cob	gob	*cab	*cog	save	shed
*code	goad	*cad	*cope	nun	chip
*core	gore	care	cone	type	want
*curl	girl	Carl	*curb	nest	soft

A.3 Experiment 3A Stimuli

Table 28: E3A: /p/-initial target stimuli.

Target	Place Neighbor	Voiced Neighbor	FillerA	Filler B
*pall	tall	ball	nub	thin
*pun	ton	bun	daft	heart
*palm	calm	balm	lard	ship
*pad	tad	bad	wool	*cool
*palm	calm	balm	thug	wren
*poll	coal	bowl	mace	yam
*pare	care	bear	yelp	jam
*putt	cut	but	nose	hag
*peek	teak	beak	guild	wrist
*pore	*core	bore	*tar	golf
*pill	till	bill	jug	vet

Table 29: E3A: /t/-initial target stimuli.

Target	Place Neighbor	Voiced Neighbor	FillerA	Filler B
*tyke	pike	dike	mast	board
*tore	*pore	door	yap	gong
*tab	*cab	dab	barn	worm
*tile	pile	dial	wedge	shoal
*tart	cart	dart	wet	loud
tote	coat	dote	mouse	gel
*torque	cork	dork	sheep	hum
*tan	pan	Dan	male	whale
*teal	*peal	deal	size	goon
*tick	kick	Dick	hill	mouth
*tuck	puck	duck	hood	moon

Table 30: E3A: /k/-initial target stimuli.

Target	Place Neighbor	Voiced Neighbor	FillerA	Filler B
*cuff	puff	guff	tool	vent
*curl	pearl	girl	mar	vain
*kilt	tilt	guilt	wolf	dim
*core	*pore	gore	*toast	yacht
*code	*toad	goad	reed	noon
*coo	two	goo	surf	rice
kale	*pale	gale	lace	bean
*cab	tab	gab	haze	wig
*cuss	pus	Guss	yard	thieve
*cod	pod	god	save	shed
*cape	tape	gape	nun	chip

A.4 Experiment 3B Stimuli

Table 31: E3B: /p/-initial target stimuli.

Target	Manner Neighbor	Voiced Neighbor	FillerA	Filler B
*pare	fare	bear	yelp	jam
*pox	fox	box	jug	vet
*pall	fall	ball	thug	wren
*pie	fye	buy	nub	thin
*pill	fill	bill	nose	hag
*poll	full	bull	wrist	guild
*pad	fad	bad	wool	*cool
*pig	fig	big	*tar	golf
*peat	feat	beat	mace	yam
*pore	four	bore	dart	fog
*punk	funk	bunk	daft	heart
*pun	fun	bun	lard	ship

Table 32: E3B: /t/-initial target stimuli.

Target	Manner Neighbor	Voiced Neighbor	FillerA	Filler B
*tuck	suck	duck	male	whale
*tick	sick	Dick	mouse	gel
*tan	sand	Dan	barn	worm
*teem	seem	deem	wet	loud
*tense	sense	dense	wit	beard
*Ted	said	dead	sheep	hum
*tore	sore	door	goose	half
*tank	sank	dank	mast	board
*tyke	psych	dike	cart	goal
*tame	same	dame	guile	mood
toe	sow	doe	sun	bed
*teal	seal	deal	yap	gong

Table 33: E3B: /k/-initial target stimuli.

Target	Manner Neighbor	Voiced Neighbor	FillerA	Filler B
cord	hoard	gourd	nun	chip
caulk	hawk	gawk	lace	bean
*kilt	hilt	guilt	tool	vent
cut	hut	gut	yard	thieve
*coo	who	goo	reed	noon
*curl	hurl	girl	*toast	yacht
*cuff	huff	guff	wolf	dim
kale	hale	gale	surf	rice
*kit	hit	git	mar	vain
call	hall	gaul	save	shed
cold	hold	gold	haze	wig
could	hood	good	nun	chip

A.5 Experiment 4 Stimuli

Table 34: E4: /p/-initial target stimuli.

Target	Voiced Neighbor	Nasal Neighbor	FillerA	Filler B
*pall	ball	mall	nub	thin
pike	bike	mike	dart	fog
*pig	big	mig	daft	heart
*punk	bunk	monk	lard	ship
*punch	bunch	munch	wool	*cool
*peat	beat	meat	thug	wren
*peek	beak	meek	mace	yam
*poll	bowl	mole	yelp	jam
*palm	balm	mom	nose	hag
*pare	bare	mare	jug	vet
*pill	bill	mill	hair	food
*pore	bore	more	shelf	mile
*putt	butt	mutt	zone	land
*pie	buy	my	nub	thin
patch	batch	match	*tar	golf
pet	bet	met	jug	vet
pelt	belt	melt	guild	wrist

Table 35: E4: /t/-initial target stimuli.

Target	Voiced Neighbor	Nasal Neighbor	FillerA	Filler B
*tile	dial	Nile	yap	gong
*tab	dab	nab	wedge	shoal
*tick	Dick	nick	size	goon
ton	done	none	hill	mouth
tip	dip	nip	hood	moon
*tame	dame	name	bench	sign
tote	dote	note	rug	shaft
*Ted	dead	Ned	cart	goal
*tore	door	nor	goose	half
*teal	deal	Neal	goof	yap
toe	doe	know	sun	bed
tale	dale	nail	mast	board
tub	dub	nub	barn	worm
tune	dune	noon	wet	loud
town	down	noun	mouse	gel
tear	deer	near	sheep	hum

Table 36: E4: /k/-initial target stimuli.

Target	Voiced Neighbor	Null-onset Neighbor	FillerA	Filler B
*cod	God	odd	tool	vent
*cuss	Guss	us	mar	vain
cap	gap	app	wolf	dim
*cab	gab	ab	surf	rice
*cape	gape	ape	yard	thieve
*code	goad	ode	nun	chip
*core	gore	ore	type	want
*curl	girl	Earl	nest	soft
kale	gale	ale	lace	bean
cord	gourd	oared	nun	chip
call	gaul	all	save	shed
cold	gold	old	haze	wig
cash	gash	ash	*toast	yacht
kill	gill	ill	reed	noon
coat	goat	oat	haze	wig
came	game	aim	save	shed
cage	gauge	age	lace	bean
cause	gauze	awes	nun	chip

A.6 Experiment 5 Stimuli

Table 37: E5: /p/-initial target stimuli.

Target	Onset Neighbor	Coda Neighbor	FillerA	Filler B
peep	keep	*peat	dart	fog
pope	*cope	poke	wrist	guess
pit	*kit	pick	hair	food
*pad	bad	pat	nub	thin
pack	tack	pat	sheep	hum
*pup	cup	*putt	*tar	golf
pat	bat	*pad	wool	*cuss
*putt	cut	puck	yelp	jam
puck	*tuck	*putt	zone	land
pot	cot	pop	size	goon
pick	*tick	pit	lard	ship
poke	coke	pope	jug	vet
*pig	big	pick	save	shed
peg	beg	peck	rug	shaft
pop	top	pot	goof	vain

Table 38: E5: /t/-initial target stimuli.

Target	Onset Neighbor	Coda Neighbor	FillerA	Filler B
tug	dug	*tuck	mace	yam
tap	cap	tack	jug	vet
tight	*kite	*tyke	yap	gong
tack	pack	tap	cart	goal
tub	*pub	tug	bench	sign
type	*pipe	*tyke	goose	half
taught	caught	talk	wedge	shoal
*tick	kick	tip	sun	bed
*tuck	duck	tug	mast	board
*tab	dab	tap	barn	worm
take	*cake	tape	mouse	gel
tape	*cape	take	hill	mouth

Table 39: E5: /k/-initial target stimuli.

Target	Onset Neighbor	Coda Neighbor	FillerA	Filler B
*cape	tape	*cake	surf	rice
coke	poke	coat	yard	thief
*cod	god	cot	nun	chip
*cab	gab	cap	wolf	dim
coat	goat	*code	lace	bean
*cake	take	*cape	reed	noon
*cop	top	cot	haze	wig
cat	pat	cap	thug	wren
*cope	pope	coke	tool	vent
cut	*putt	cup	lace	bill
kick	*tick	*kit	hood	moon
cot	got	*cod	shelf	mile
*kit	pit	kick	nose	hall
cup	*pup	cut	daft	heart
cap	gap	*cab	*toast	yacht

A.7 Experiment 6 Stimuli

Table 40: E6: /p/-initial target stimuli.

Target	Voiced Neighbor	FillerA	Filler B
pack	back	guild	wrist
*pad	bad	norm	van
*pall	ball	nub	thin
*palm	balm	nose	hag
*pare	bare	jug	vet
pat	bat	wool	gown
patch	batch	*tar	golf
*peat	beat	thug	wren
*peek	beak	mace	yam
peg	beg	rug	shaft
pelt	belt	chair	gum
pet	bet	gun	will
*pie	buy	wreck	roof
*pig	big	daft	heart
pike	bike	dart	fog
*pill	bill	hair	food
*poll	bowl	yelp	jam
*pore	bore	shelf	mile
*pox	box	gull	sheath
*pun	bun	sieve	hole
*punch	bunch	lag	*cool
*punk	bunk	lard	ship
*putt	butt	zone	land

Table 41: E6: /t/-initial target stimuli.

Target	Voiced Neighbor	FillerA	Filler B
*tab	dab	wedge	shoal
tale	dale	mast	board
*tame	dame	bench	sign
*tan	Dan	sheep	hum
*tank	dank	mesh	rogue
*tart	dart	mouse	gel
*taunt	daunt	guile	mood
*teal	deal	goof	yap
*teem	deem	bog	gush
*tense	dense	wit	beard
tent	dent	barn	worm
*tile	dial	hem	gong
tip	dip	hood	moon
toe	doe	sun	bed
*tomb	doom	male	whale
ton	done	hill	mouth
*tore	door	goose	half
*torque	dork	wet	loud
tote	dote	gene	chief
town	down	shake	hell
tub	dub	mall	chalk
*tuck	duck	meal	bone
tug	dug	mine	fib
*tyke	dike	lab	gem
tot	dot	fold	road

Table 42: E6: /k/-initial target stimuli.

Target	Voiced Neighbor	FillerA	Filler B
*cab	gab	surf	rice
cage	gauge	lace	bean
call	gaul	save	shed
came	game	chirp	wreath
cap	gap	wolf	dim
*cape	gape	yard	thieve
cash	gash	*toast	yacht
caulk	gawk	rod	dish
cause	gauze	nun	chip
coat	goat	haze	wig
*cob	gob	hitch	lore
*cod	god	tool	vent
*code	goad	daze	lure
cold	gold	far	lurk
*coo	goo	fat	jack
cord	gourd	shack	hutch
*core	gore	type	want
cot	got	chain	zing
could	good	wear	join
*cuff	guff	reed	noon
*curl	girl	nest	soft
cut	gut	din	knight
kale	gale	hose	sub
kill	gill	nail	math
*kilt	guilt	niche	womb
*kit	git	full	shove

B Phonological Features

This dissertation assumes the following binary feature set:

- Consonantal (cons)
- Approximant (approx)
- Sonorant (son)
- Continuant (cont)
- Nasal (nas)
- Voiced (voice)
- Spread Glottis (spread)

- Labial (lab)
- Coronal (cor)
- Anterior (ant)
- Strident (strid)
- Lateral (lat)
- Dorsal (dors)
- High (high)
- Back (back)
- Tense (tense)
- Mid (mid)
- Central (central)

Table 43: Consonant Features

	p	t	ʧ	k	b	d	ʤ	g	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	l	r	j	w
cons	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
approx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
son	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
cont	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
nas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
voice	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
spread	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
lab	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
cor	0	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	1	0	1	0	1	1	0	0
ant	0	1	0	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	1	0	1	0	0	0
strid	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
lat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
dors	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
high	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
tense	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
central	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 44: Vowel Features

	æ	ɑ	ɔ	ɛ	e	ɪ	i	o	ʊ	u	ə	ʌ
cons	0	0	0	0	0	0	0	0	0	0	0	0
approx	0	0	0	0	0	0	0	0	0	0	0	0
son	1	1	1	1	1	1	1	1	1	1	1	1
cont	1	1	1	1	1	1	1	1	1	1	1	1
nas	0	0	0	0	0	0	0	0	0	0	0	0
voice	1	1	1	1	1	1	1	1	1	1	1	1
spread	0	0	0	0	0	0	0	0	0	0	0	0
lab	0	0	0	0	0	0	0	0	0	0	0	0
cor	0	0	0	0	0	0	0	0	0	0	0	0
ant	0	0	0	0	0	0	0	0	0	0	0	0
strid	0	0	0	0	0	0	0	0	0	0	0	0
lat	0	0	0	0	0	0	0	0	0	0	0	0
dors	0	0	0	0	0	0	0	0	0	0	0	0
high	0	0	0	0	0	1	1	0	1	1	0	0
back	0	1	1	0	0	0	0	1	1	1	0	0
tense	0	1	0	0	1	0	1	1	0	1	0	1
mid	0	0	1	1	1	0	0	1	0	0	0	0
central	0	0	0	0	0	0	0	0	0	0	1	1

Bibliography

Alario, F.-X. & Moscoso Del Prado Martín, F. (2010). On the origin of the cumulative semantic inhibition effect. *Memory and Cognition*, *38*(1), 57–66.

Amano-Kusumoto, A. & Hosom, J.-P. (2011). A review of research on speech intelligibility and correlations with acoustic features. Technical Report CSLU-011-001, Oregon Health & Science University.

Ardi, R. (2004). Seriality of phonological encoding in naming objects and their names. *Memory and Cognition*, *32*(2), 212–222.

Ardi, R. & Piai, V. (2013). Associative facilitation in the stroop task: Comment on Mahon et al. (2012). *Cortex*, *49*(1), 1767–1769.

Aylett, M. & Turk, A. (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.

Baayen, R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Linguistics*. Cambridge, UK: Cambridge University Press.

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

Bailey, T. M. & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*(3), 339–362.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Beattie, G. W. & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211.

Beckman, J., Helgason, P., McMurray, B., & Ringen, C. (2011). Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of Phonetics*, *39*(1), 39–49.

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

Benkí, J. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica*, *60*(2), 129–157.

Bishop, C. M. & Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*, volume 1. New York: Springer.

Bonneau, A. (1996). Identification of vowel features from French stop bursts. In *Proceedings of the 4th International Conference on Spoken Language Processing*, (pp. 2506–2509)., Philadelphia, PA.

Bowdle, B. F. & Wright, R. (1998). Lexical neighborhoods and subjective intelligibility ratings: A preliminary report. Technical report, Indiana University.

Bradlow, A. R. (2002). Clonfluent talker- and listener-oriented forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Papers in Laboratory Phonology VII* (pp. 241–273). Mouton de Gruyter.

Bradlow, A. R. & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recongnition by native and non-native listeners. *Journal of the Acoustical Society of America*, *121*(4), 2339–2349.

Bradlow, A. R., Torretta, G., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3), 255–272.

Buz, E. & Jaeger, T. F. (2012). Effects of phonological confusability on speech duration. Poster presented at the 25th City University of New York Sentence Processing Conference. New York, NY.

Chen, Q. & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430.

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

Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyperarticulation in Korean. *Journal of Phonetics*, *39*(3), 344–361.

Clopper, C. G. & Pierrehumbert, J. B. (2008). Effects of semantic predictability and regional dialect on vowel space reduction. *Journal of the Acoustical Society of America*, *124*(3), 1682–1688.

Cohen-Priva, U. & Jurafsky, D. (2008). Phone information content influences phone duration. Poster presented at the Conference on Prosody and Language Processing. Cornell University.

Cooke, M. & Lu, Y. (2009). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *Journal of the Acoustical Society of America*.

Costa, A. & Sebastian-Gallés (1998). Abstract phonological structure in language production: Evidence from Spanish. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 886–903.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley and Sons.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language*, *54*(1), 1–19.

Creel, S. C., Delphine, D., & Swingley, D. (2006). Effects of featural similarity and overlap position on lexical confusions and overt similarity judgements. In *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburg, PA.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations. *Cognition*, *108*(3), 710–718.

Damian, M. F. & Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*(3), B77–B78.

Davis, C. J. & Taft, M. (2005). More words in the neighborhood: Interference in lexical decision due to deletion neighbors. *Psychonomic Bulletin and Review*, *12*(5), 904–910.

De Cara, B. & Goswami, U. (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods, Instruments & Computers*, *34*(3), 416–423.

de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics*, *32*(4), 493–516.

de Jong, K. & Zawaydeh, B. (2002). Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration. *Journal of Phonetics*, *30*(1), 53–75.

de Zubicaray, G. I., Hansen, S., & McMahon, K. L. (2013). Differential processing of thematic and categorical conceptual relations in spoken word production. *Journal of Experimental Psychology: General*, *142*(1), 131–142.

de Zubicaray, G. I., McMahon, K. L., Eastburn, M. M., & Wilson, S. J. (2002). Orthographic/phonological facilitation of naming responses in the picture-word task: An event-related fMRI study using overt vocal responsing. *NeuroImage*, *16*(4), 1084–1093.

de Zubicaray, G. I., Miozzo, M., Johnson, K., Schiller, N. O., & McMahon, K. L. (2012). Independent distractor frequency and age-ofacquisition effects in picture-word interference: fMRI evidence for post-lexical and lexical accounts according to distractor type. *Journal of Cognitive Neuroscience*, *24*(2), 482–495.

Dell, F. & Elmedlaoui, M. (1985). Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics*, *7*, 105–130.

Dell, G. S. & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 9–39). Mouton, New York.

Demberg, V., Sayeed, A. B., Gorinski, P. J., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 356–367)., Jeju, Korea. Association for Computational Linguistics.

Dilts, P., Baayen, R. H., & Tucker, B. V. (2011). Word duration and segment deletion as measures of reduction in a corpus of spontaneous speech. *Journal of the Acoustical Society of America*, *129*(4), 2455.

Erlhagen, W. & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Everett, C., Miller, Z., Nelson, K., Soare, V., & Vinson, J. (2011). Reduction of Brazilian Portuguese vowels in semantically predictable contexts. In *Proceedings of the 17th International Congress of Phonetic Sciences*, (pp. 548–551)., Hong Kong, China.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Feldman, N. H., Morgan, J. L., & Griffiths, T. L. (2009). The influence of categories on perception: Explainng the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Flemming, E. (2010). Modeling listeners: Comments on Pluymaekers et al. and Scarborough. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory Phonology*, volume 10 (pp. 587–606). Berlin: Mouton de Gruyter.

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

Fowler, C. A. (1988). Differential shortening of repeated content words reduced in various communicatie contexts. *Language and Speech*, *31*(4), 307–319.

Frank, A. F. & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, (pp. 939–944)., Washington, D.C.

Gahl, S. (2012a). Phonetic reduction in the "easy/hard" database without "difficulty". Technical report, University of California, Berkeley.

Gahl, S. (2012b). Why so short? competiting explanations for variation. In *In Proceedings of the 30th West Coast Conference on Formal Linguistics*, Santa Cruz, CA.

Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, *71*(1), 1138–1149.

Genzel, D. & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (pp. 199–206)., Philadelphia, PA.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–934.

Goldberg, A. (2010). *Gradient well-formedness across the morpheme boundary*. PhD thesis, Johns Hopkins University.

Goldinger, S. & Summers, W. V. (1989). Lexical neighborhoods in speech production: A first report. In *Research on Speech Percetion Progress Report*, number 15, (pp. 331–342)., Bloomington.

Goldrick, M. (2008). A gradient harmonic grammar account of lexically-conditioned phonetic variation. Talk presented at the 82nd Annual Meeting of the Linguistic Society of America. Chicago, IL.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*(2), 113–134.

Goldrick, M. & Rapp, B. (2002). A restricted interaction account RIA of spoken word production: The best of both worlds. *Aphasiology*, *16*(1/2), 20–55.

Gordon, P. C. & Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 153–178.

Griffin, Z. M. & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*(3), 313–338.

Griffin, Z. M. & Ferreira, V. S. (2006). Properties of spoken language production. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics, 2nd Edition* (pp. 21–59). London, UK: Academic Press.

Guenther, F., Hampson, M., & Johnson, D. (1997). A theoretical investigation of reference frames for the planning of speech movements. Technical Report CAS/CNS-97-002, Boston University.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*(3), 594–621.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, volume 2, (pp. 159–166).

Harley, T. A. (2001). *The Psychology of Language: From Data to Theory*. Hove: Psychology Press.

Hayes, B. & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Hazan, V. (2012). The effect of speaker-listener interaction on speech production in adverse listening conditions. Talk presented at the Listening Talker Workshop. Edinburgh.

Hazan, V. & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America*, *130*(4), 2139–2152.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Hilliard, C., Furth, K., & Jaeger, F. (2011). Phonological encoding in sentence production. In *Proceeding of the 33rd Annual Meeting of the Cognitive Science Society*, (pp. 3070–3075)., Boston, MA.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: experimental and computational studies. *Cognition*, *100*(3), 464–482.

Jaeger, F. T., Furth, K., & Hilliard, C. (2012a). Phonological encoding during unscripted sentence production. *Frontiers in Psychology*, *3*(481).

Jaeger, F. T., Furth, K., & Hilliard, C. (2012b). Phonological overlap affects lexical selection during sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1439–1449.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*(230).

Jaeger, T. F. & Ferreira, V. (2013). Seeking predictions from a predictive framework. *Behavioral and Brain Sciences*, *36*(4), 31–32.

Jaynes, E. T. (1968). Prior probabilities. *IEEE on Systems Science and Cybernetics*, *4*(3), 227–241.

Jescheniak, J. D., Schriefers, H., & Hantsch, A. (2003). Utterance formant affects phonological priming in the picture-word task: Implications for models of phonological encoding in speech

production. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 441–454.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, *69*(3), 505–528.

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 317–324)., Barcelona.

Kessler, B., Treiman, R., & Mullenix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, *47*(1), 145–171.

Kinoshita, S. & Norris, D. (2009). Transposed-letter priming of prelexical orthographic representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 1–18.

Kirby, J. P. & Yu, A. C. (2009). Morphological paradigm effects on phonetic realization. University of Chicago ms.

Kirov, C. & Wilson, C. (2012). The specificity of online variation in speech production. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, (pp. 587–592)., Sapporo, Japan.

Kleinman, D. (2013). Resolving semantic interference during word production requires central attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1860–1877.

Knill, D. & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge, UK: Cambridge University Press.

Kolne, K. (2011). The effect of phonological similarity and word length on rapid naming performance. Master's thesis, McMaster University.

Kraljic, T. & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kuperman, V. & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611.

Lam, T. Q. (2012). *The Prominence of Referring Expressions: Message and Lexical Level Effects*. PhD thesis, University of Illinois.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933.

Lau, P. (2008). The Lombard Effect as a communicative phenomenon. Technical report, University of California Berkeley.

Lefkowitz, M. L. (2012). The nature of phonetic disassociation from lexical neighbors. Master's thesis, University of California, Los Angeles.

Lehnert-LeHouillier, H. (2010). Investigating lexically conditioned phonetic variation with ultrasound. Talk presented at Ultrafest. New Haven, CT.

Levelt, W. J. (1993). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. (1999). Models of word production. *Trends in Cognitive Science*, *3*(6), 223–232.

Levelt, W. J., Ardi, R., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75.

Levy, R. (2005). *Probabilistic Models of Word Order and Syntactic Discontinuity*. PhD thesis, Stanford University.

Levy, R. & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J., & Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems*, volume 19, (pp. 849–856)., Cambridge, MA. MIT Press.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. Hardcastle & A. Maschal (Eds.), *Speech Production and Speech Modeling* (pp. 403–439). Kluwer Academic Publishers.

Lisker, L. & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Conference of Phonetic Sciences*, (pp. 563–567)., Cambridge. Cambridge University Press.

Luce, P. A. (1986). *Neighborhoods of Words in the Mental Lexicon*. PhD thesis, Indiana University.

Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activativation model. *Ear and Hearing*, *19*(1), 1–36.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, *28*(2), 203–208.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*(226).

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order or speech. *Neuropsychologia*, *8*(3), 323–350.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Mädebach, A., Opperman, F., Hantsch, A., Curda, C., & Jescheniak, J. D. (2011). Is there semantic interference in delayed naming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 522–538.

Madsen, R. E., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning*, (pp. 545–552)., Bonn, Germany. Association for Computing Machinery.

Maess, B., Friederici, A. D., Damian, M., Meyer, A. S., & Levelt, W. J. (2002). Semantic category interference in overt picture naming: Sharpening current density localization by PCA. *Journal of Cognitive Neuroscience*, *14*(3), 455–462.

Mahon, B. Z., Peterson, R., Costa, A., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503–535.

Manin, D. Y. (2006). Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes*, *6*(3), 229–236.

Marcel, A. J., Suslick, R. L., Michaels, S., & Shockey, L. (1978). Acoustic cues and psychological processes in the perception of natural stop consonants. *Perception & Psychophysics*, *24*(4), 327–336.

Mark, S. & Wheeldon, L. (2004). Horizontal information flow in spoken sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 675–686.

Marslen-Wilson, W. & Zwitserlood, P. (1989). Accessing spoken words: The importance of word

onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, *2*(6), 387–395.

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, *29*(5), 524–545.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language*, *30*(1), 69–89.

Meyer, A. S., Ardi, R., & Levelt, W. J. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, *48*(1), 131–147.

Meyer, A. S. & Belke, E. (2007). Word form retrieval in language production. In G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 471–487). Oxford, UK: Oxford University Press.

Meyer, A. S. & Van Der Meulen, F. F. (2000). Phonological priming effects on speech onset latencies and viewing times in object naming. *Psychonomic Bulletin and Review*, *7*(2), 314–319.

Meyer, D. E. & Gordon, P. C. (1985). Speech production: Motor programming of phonetic features. *Journal of Memory and Language*, *24*(1), 3–26.

Mirman, D. (2011). Effects of near and distant semantic neighbors on word production. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(1), 32–43.

Mirman, D., Kittredge, A. K., & Dell, G. S. (2010). Effects of near and distant phonological neighbors on picture naming. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, (pp. 1447–1452)., Portland, OR.

Moat, H. S. (2010). *Modelling subphonemic information flow: An investigation and extension of Dell's (1986) model of word production*. PhD thesis, University of Edinburgh.

Mooshammer, C. R., Goldstein, L., Tiede, M., Kulshreshtha, M., McClure, S., & Katsika, A. (2009). Planning time effects of phonological competition: articulatory and acoustic data. *The Journal of the Acoustical Society of America*, *125*(4), 2657.

Morsella, E. & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 555–563.

Mulatti, C., Peressotti, F., Job, R., Saunders, S., & Coltheart, M. (2012). Reading aloud: The cumulative lexical interference effect. *Psychonomic Bulletin and Review*, *19*(4), 662–667.

Munson, B. (2013). The influence of production latencies and phonological neighborhood density on vowel dispersion. *Journal of the Acoustical Society of America*, *133*(5), 3567.

Munson, B. & Babel, M. E. (2005). The sequential cueing effect in children's speech production. *Applied Psycholinguistics*, *26*(2), 157–174.

Munson, B. & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1048–1058.

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*, *51*(3), 574–585.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Norris, D. & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Perea, M. & Lupker, S. J. (2003). Transposed-letter confusability effects in masked form priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked Priming: State of the Art* (pp. 97–120). Hove, UK: Psychology Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, (pp. 2582–2587)., Amsterdam.

Plotkin, J. B. & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology*, *205*(147-159).

Port, R. F. & Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America*, *66*(3), 654–662.

Qian, T. & Jaeger, T. F. (2010). Entropy profiles in language: A cross-linguistic investigation. Manuscript submitted for publication.

Rahman, R. A. & Aristei, S. (2010). Now you see it... and now again: Semantic interference reflects lexical competition in speech production with and without articulation. *Psychonomic Bulletin and Review*, *17*(5), 657–661.

Rahman, R. A. & Sommer, W. (2003). Does phonological encoding in speech production always follow the retrieval of semantic knowledge? Electrophysical evidence for parallel processing. *Cognitive Brain Research*, *16*(3), 372–382.

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review*, *120*(1), 281–292.

Repp, B. H. & Lin, H.-B. (1988). Acoustic properties and perception of stop consonant release transients. Technical Report SR-95/96, Haskins Laboratories.

Rey, A., Courrieu, P., Madec, S., & Grainger, J. (2013). The unbearable articulatory nature of naming: on the reliability of word naming responses at the item level. *Psychonomic Bulletin and Review*, *20*(1), 87–94.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*(3), 249–284.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, *14*(2), 173–200.

Roelofs, A. & Piai, V. (2011). Attention demands of spoken word planning: A review. *Frontiers in Psychology*, *2*(307).

Rogers, M. A. & Storkel, H. L. (1998). Reprogramming phonologically similar utterances: The role of phonetic features in pre-motor encoding. *Journal of Speech, Language, and Hearing Research*, *41*(2), 258–274.

Roon, K. (2012). *The Dynamics of Phonological Planning*. PhD thesis, New York University.

Roon, K. D. & Gafos, A. I. (2013). A dynamical model of the speech perception-production link. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, (pp. 1241–1246)., Berlin, Germany.

Sadat, J., Martin, C., Costa, A., & Alario, F.-X. (2012). Phonological neighbourhood in speech production revisited. Poster presented at the 7th International Workshop on Language Production. New York University.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., & Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within broca's area. *Science*, *326*(5951), 445–449.

Saltzman, E. & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Santesteban, M., Costa, A., Pontin, S., & Navarrete, E. (2006). The effect of word-frequency on lexical selection in speech production: Evidence from semantic homogeneous naming contexts. *Cognitiva*, *18*(1), 75–84.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17.

Scarborough, R., Styler, W., & Zellou, G. (2011). Nasal coarticulation in lexical perception: The role of neighborhood-conditioned variation. In *Proceedings of the 17th International Congress of Phonetic Sciences*, (pp. 1750–1753)., Hong Kong, China.

Scarborough, R. & Zellou, G. (2012). Continua of clarity: Lexical neighborhoods and clear speech. Poster presented at the 13th Conference on Laboratory Phonology. Stuttgart, Germany.

Scarborough, R. A. (2004). *Coarticulation Structure and the Lexicon*. PhD thesis, University of California, Los Angeles.

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics*, *41*(3), 249–263.

Schnur, T. T. (2011). Phonological planning during sentence production: beyond the verb. *Frontiers in Psychology*, *2*(319).

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199–227.

Sevald, C. A. & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, *53*(2), 91–127.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3-4), 379–423,623–656.

Shaw, J. A. (2012). Metrical rhythm in speech planning: priming or predictability. In *Proceedings of the Conference on Speech Science and Technology*, (pp. 145–148)., Sydney, Australia. Australian Speech Science and Technology Association.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Siegelman, H. T. & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, *50*(1), 132–150.

Simoncelli, E. P. (2009). Optimal estimation in sensory systems. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences*. MIT Press.

Smiljanić, R. & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America*, *118*(3), 1677–1688.

Smiljanić, R. & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, *3*(1), 236–264.

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2 (pp. 390–431). MIT Press.

Smolensky, P. & Legendre, G. (2006). *The Harmonic Mind*, volume 2. The MIT Press.

Spalek, K. & Damian, M. F. (2013). Picture-word interference with masked and visible distractors: Different types of semantic relatedness inhibit lexical selection. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Societymeeting of the Cognitive Science Society*, (pp. 1366–1371)., Berlin, Germany.

Spencer, K. A. & Wiley, E. (2008). Response priming patterns differ with interstimulus interval duration. *Clinical Linuistics and Phonetics*, *22*(6), 475–490.

Stocker, A. A. & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Suchato, A. & Punyabukkana, P. (2005). Factors in classification of stop consonant place of articulation. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, (pp. 2969–2972)., Lisbon, Portugal.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, *84*(3), 917–928.

Surendran, D. & Niyogi, P. (2006). Quantifying the functional load of phonemic positions, distinctive features, and suprasegmentals. In O. N. Thomsen (Ed.), *Competing Models of Linguistic Change: Evolution and Beyond*, volume 279. Amsterdam: John Benjamins.

Teodorescu, A. R. & Usher, M. (2013). Disentangling decision models: Fromt independence to competition. *Psychological Review*, *120*(1), 1–38.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2007). The effect of speaking rate on voice-onset-time is talker-specific. In *Proceedings of the 16th Annual International Congress of Phonetic Sciences*, (pp. 473–476)., Saarbrücken, Germany.

Tilsen, S. (2007). Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming. Technical report, UC Berkeley.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, *37*(3), 276–296.

Tilsen, S. (2013). Inhibitory mechanisms in speech planning maintain and maximize contrast. In A. Yu (Ed.), *Origins of Sound Patterns: Approaches to Phonologization*. Oxford, UK: Oxford University Press.

Tilsen, S. & Goldstein, L. (2012). Articulatory gestures are individually selected in production. *Journal of Phonetics*, *40*(6), 764–779.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2), 147–165.

Treiman, R., Clifton Jr., C., Meyer, A. S., & Wurm, L. H. (2003). Language comprehension and production. In *Comprehensive Handbook of Psychology, Volume 4: Experimental Psychology* (pp. 527–548). New York: John Wiley and Sons.

Uchanski, R. M. (2008). Clear speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 207–235). Oxford, UK: Blackwell.

van Son, R., Bolotova, O., Lennes, M., & Pols, L. C. (2004). Frequency effects on vowel reduction in three typologically different languages (Dutch, Finnish, Russian). In *Proceedings of the International Conference on Spoken Language Processing*, (pp. 1277–1280)., Jeju, Korea.

van Son, R. & Pols, L. C. (2002). Evidence for efficiency in vowel production. In *Proceedings of the International Conference on Spoken Language Processing*, (pp. 37–40)., Denver, CO.

van Son, R. & Pols, L. C. (2003). How efficient is speech? In *Proceedings of the 15th International Congress of Phonetic Sciences*, (pp. 171–184)., Barcelona.

VanDam, M. & Silbert, N. (2010). A Bayesian model of voice-onset time production. *Journal of the Acoustical Society of America*, *127*(3), 1853–1853.

Vitevitch, M. S. (2002a). Influence of onset density on spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 270–278.

Vitevitch, M. S. (2002b). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735–747.

Vitevitch, M. S., Armbrüster, J., & Chu, S. (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. & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Vitevitch, M. S. & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 481–487.

Vitevitch, M. S. & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, *52*(2), 193–204.

Vitevitch, M. S. & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*(6), 760–770.

Vitkovitch, M. & Cooper, E. (2012). My word! interference from reading object names implies a role for competition during picture name retrieval. *The Quarterly Journal of Experimental Psychology*, *65*(6), 1229–1240.

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech*, *56*(3), 395–417.

Welby, P. (2006). Intonational differences in lombard speech: Looking beyond f_0 range. In Hoffman, R. & Mixdorff, H. (Eds.), *Proceedings of the 3rd International Conference on Speech Prosody*, (pp. 763–766)., Dresden, Germany.

Wheeldon, L. (2003). Inhibitory form priming of spoken word production. *Language and Cognitive Processes*, *18*(1), 81–109.

Wiener, S., Speer, S. R., & Shank, C. (2012). Effects of frequency, repetition and prosodic location on ambiguous mandarin word production. In *Proceedings of the 6th International Conference on Speech Prosody*, (pp. 528–531)., Shanghai, China.

Wright, R. (2003). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI* (pp. 75–87). Cambridge University Press.

Yaniv, I., Meyer, D. E., Gordon, P. C., Huff, C. A., & Sevald, C. A. (1990). Vowel similarity, connectionist models, and syllable structure in motor programming of speech. *Journal of Memory and Language*, *29*(1), 1–26.

Yao, Y. (2010). Separating speaker and listener-oriented forces in speech: Evidence from phonological neighborhood density. Talk presented and the 84th Annual Meeting of the Linguistic Society of America. Baltimore, MD.

Yap, T. F., Epps, J., Ambikairajah, E., & Choi, E. H. C. (2011). Formant frequencies under cognitive load: Effects and classification. *EURASIP Journal on Advances in Signal Processing*, *1*(219253).

Zhang, S., Lee, M. D., Vandekerckhove, J., Maris, G., & Wagenmakers, E.-J. (2012). On the relationhip between diffusion and accumulator sequential sampling models. Manuscript submitted for publication.

Zhao, Y. & Jurafsky, D. (2009). The effect of lexical frequency and Lombard Reflex on tone hyperarticulation. *Journal of Phonetics*, *37*(2), 231–247.

Vita

Christo Kirov was born on May 24, 1984 in Sofia, Bulgaria. He received a BA in Linguistics and Computer Science from New York University in May of 2007, and enrolled in the Ph.D. program in Cognitive Science at Johns Hopkins University in September of 2007. His research focuses on understanding speech production and perception using empirical and formal methods. He has presented his work at a number of conferences in linguistics and cognitive science, and has been published in major journals and conference proceedings.