Compression effects, perceptual asymmetries, and the grammar of timing

by

Jonah Katz

B.A. Linguistics, Music
University of Massachusetts Amherst, 2003

SUBMITTED TO THE DEPARTMENT OF LINGUISTICS AND PHILOSOPHY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN LINGUISTICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2010

Signature of Author: _____
Department of Linguistics and Philosophy
September 7, 2010

Certified by: _____
Edward Flemming
Associate Professor of Linguistics
Thesis Supervisor

_____
Donca Steriade
Class of 1941 Professor of Linguistics
Thesis Committee Member

_____
Adam Albright
Associate Professor of Linguistics
Thesis Committee Member

Accepted by: _____
Irene Heim
Professor of Linguistics
Department Head

Compression effects, perceptual asymmetries, and the grammar of timing

by

Jonah Katz

Submitted to the Department of Linguistics & Philosophy
on September 10, 2010 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Linguistics

ABSTRACT

This dissertation reports the results of two English experiments on timing and perception. The first experiment demonstrates asymmetries in timing between consonants and vowels, which depend on the manner of the consonant. The second experiment shows that these asymmetries in speech production are mirrored by perceptual asymmetries among consonants with different manner features. We argue that these phenomena are best described in terms of auditory rather than articulatory representations. A formal analysis is developed using weighted, gradiently-violable constraints on segment and syllable duration. Because the constraints make reference to the auditory features of segments, the analysis can derive the relationship between asymmetries in speech production and asymmetries in speech perception. The patterns of timing discovered here appear to interact in limited ways with systems of phonological contrast. We incorporate the duration constraints proposed here into a phonetically-driven model of phonology, examining the predictions that such an approach makes about phonological typology.

Thesis Supervisor: Edward Flemming
Title: Associate Professor of Linguistics

Acknowledgements

A large number of colleagues, teachers, and friends have helped me get to this point, and I can't possibly thank them all here. Nonetheless, I would like to single out a few people who have been especially important to me over the preceding years.

Edward Flemming is an extraordinarily knowledgeable and insightful linguist, and it is hard to imagine a better thesis advisor. At every step of this project, he has struck a balance between pushing me to take the work in new directions and making sure that what I've already done is up to the same standards of empirical and formal rigor to which he holds his own work. Nothing contained here would have been possible without Edward's help; he supervised this work from the day in my second year that I decided to try an experiment on compensatory shortening to the day that I completed my revisions. Edward is an inspiring teacher, mentor, and role model.

Adam Albright has been instrumental in the development of this thesis. No matter how abstract the theoretical idea one presents to him, he always seems to have a range of relevant phonological data points at his fingertips. He constantly pushes his students to broaden the empirical coverage of their work to account for patterns that they may not have known about or taken to be in the purview of their theories. Adam is also the only teacher I've ever had who quotes the Simpsons as much as I do, and that is worth something in and of itself.

Donca Steriade is an inspiration to everybody who works with her. She has the uncanny property of always being two steps ahead of you, even when you're describing your own theory to her. She has constantly pushed me to pay attention to the details, strive for formal precision and clarity, and think through the broader consequences of every hypothesis. Donca has constantly challenged me to sharpen and broaden my thinking, with complete confidence in my ability to rise to those challenges.

Many other people have offered helpful comments on various parts of this project. Thanks to Lasse Bombien, Adamantios Gafos, Jelena Krivokapic, Elliott Moreton, Joe Pater, Jason Shaw, Mark Tiede, and audiences at MUMM 2007 and the Munich Workshop on Consonant Clusters and Structural Complexity.

Lisa Selkirk has done a few things for me over the years. She taught my first phonology class, designed my first experiment, gave me my first research assistantship, convinced me to go to grad school, co-authored my first paper, and helped me get my first job. Lisa is the *sine qua non* of my linguistic life; I'm terribly fortunate to have met her, and she is an unbelievable teacher, mentor, colleague, and friend.

David Pesetsky has been very important to me, as a linguist and as a friend. He nearly convinced me to study syntax during my first few years at MIT. Even though that didn't end up happening, we've remained very close. David taught me how to form an argument, and how to write (and think) clearly. Whether we are writing a paper together, sparring with musicologists in Berlin, or discussing how to design a text-to-speech system with a thick Russian accent, David is a joy to be around. Even in an environment as full of wonderful teachers as the one I've come from, he stands out for his sheer dedication to the well-being and intellectual advancement of his students.

John Kingston taught the hardest undergraduate class I ever took, and it also turned out to be the most important. After I graduated, John hired me as a full-time research assistant in his lab; that's where I learned to be a scientist and to do phonetics. He has also had an important influence on the work reported here; his words of encouragement after hearing about an earlier version of the production study helped convince me that this work was worth pursuing, and suggested some directions in which I might pursue it.

All of my teachers at MIT, and as an undergraduate at UMass, have been unbelievably supportive and helpful. Many thanks to Kai von Fintel, Danny Fox, Irene Heim, Sabine Iatridou, Kyle Johnson, Eva Juarros-Daussà, Michael Kenstowicz, Angelika Kratzer, Shigeru Miyagawa, Barbara Partee, and Norvin Richards.

I've learned from many other students at MIT during my time here, and benefited from their friendship. I single out my amazing classmates Hyesun Cho, Jessica Coon, Gillian Gallagher, Maria Giavazzi, and Patrick Jones for special gratitude.

My friends, my family, and especially my wonderful parents Larry and Stephanie have offered unconditional love and support throughout my grad school experience, and I doubt I could have done it without them. Jodie Rose, in particular, has put up with a very boring Jonah for the last year or so, and managed to remain unfailingly upbeat, supportive, and fun the entire time.

Table of Contents

1 Introduction

This dissertation is concerned with timing patterns in speech production, the auditory and articulatory influences on those patterns, and the way those patterns interact with phonological contrasts. We argue that some aspects of timing must be explained with reference to auditory representations rather than articulatory ones. A formalism is developed that derives timing patterns in production from auditory properties of the units to be produced. That formalism relies on assumptions about auditory perception, which are tested experimentally. Finally, the formalism is extended to account for certain categorical phonotactic phenomena.

The grammar of timing is a rather broad topic, and no single work will settle all of the questions inherent to the domain. This dissertation approaches the topic starting from a narrow range of phenomena in English. These phenomena are known as *compression* or *compensatory shortening* effects. Evidence from this domain is then incorporated into a general model of timing.

The term 'compression' refers to the fact that, when more segments are present in a syllable, each one of the segments is shorter. For instance, /æ/ in *sad* is shorter than in *add*. This holds in English for the addition of both onset (Fowler 1983, van Santen 1992) and coda consonants (Fowler 1983, Munhall *et al.* 1992). It also holds in Dutch (Waals 1999) and Swedish (Lindblom & Rapp 1973). The typologically widespread phenomenon known as *closed-syllable vowel shortening* (Maddieson 1985) is the most familiar compression effect. We further distinguish *simplex compression*, between an item with no consonant in a given position and an item with a

consonant in that position (e.g. *add-sad*); from *incremental compression*, between an item with one consonant and several (e.g. *lad-clad*).

No previous study has examined the influence of consonant manner or syllable position on compression effects. Chapter 2 reports the results of an English nonce-word study that examined obstruents, nasals, liquids, and clusters in onset and coda position. All consonants are associated with some amount of simple vowel-compression, but not all strings induce incremental compression. Clusters including liquids induce incremental compression in both onset and coda position relative to liquid singletons, clusters including nasals do so only in onset position, and clusters containing only obstruents do not condition incremental compression in either position. For instance, the vowels in /brod/ and /dɔrb/ are significantly shorter than those in /rod/ and /dɔr/, but the vowel in /donz/ is not shorter than that in /don/.

The results have broad consequences for the theory of timing. One common analysis of compression effects treats them as emergent from general principles of articulatory gestural organization (Fowler 1983, Nam *et al.* 2009). When the articulatory gestures that are associated with segments overlap more, the acoustic manifestations of those segments will be shorter. Thus, patterns of compression should correspond to independent facts about the temporal organization of gestures. The asymmetries in incremental compression reported here, however, can not be explained by any known facts about gestural organization in English. While articulatory studies find that consonant clusters impinge more on a following vowel than singleton consonants (part of a phenomenon known as the *C-center effect*), the same is not generally found for coda consonants (Browman & Goldstein 1990 *et seq.*, Honorof & Browman 1995). Even if the C-

center effect is extended to syllable codas, differences between various manners of consonant are difficult to explain in articulatory terms.

We argue instead that compression effects are due to constraints on the auditory duration of segments and syllables. This generalizes a common approach to closed-syllable vowel shortening: compression effects are due to conflicting pressures on segments and larger units such as rimes or syllables (Maddieson 1985, Fujimura 1987, Flemming 2001). Consonants behave differently with respect to compression because constraints on duration are stated in perceptual terms, and consonants differ widely in how much perceptual information they contain about an adjacent vowel.

In this approach, patterns of compression can be explained in terms of independent facts about perception. For instance, vowels shorten more adjacent to liquids than adjacent to obstruents because liquids help to satisfy the duration requirements of an adjacent vowel more than obstruents do. This, in turn, is because liquids contain more information about adjacent vowels than obstruents do. Incorporating these hypotheses into a formal grammar requires some adjustment to the notion of a segment's duration; the end result is that the grammar manipulates something more like a segment's *recoverability*.

In chapter 3, a formal theoretical model is developed to account for asymmetries in compression. Both segments and syllables have auditory duration targets; weighted constraints assess a *cost* to any linguistic form related to the difference between target durations and realized durations in that form. As segments are added into a syllable, conflict arises between the pressure to keep

segments long and the pressure to keep syllables short. The weighted-constraint formalism predicts that the result should be a compromise between the two pressures. This is exactly what was found in the chapter 2.

Duration constraints on segments assess not only the duration of the segment itself, but the duration and amount of perceptual information included in adjacent transitions and segments. The model thus predicts that, in cases where two segments or their transitions differ in these properties, they should also differ with regard to compression effects. We develop several hypotheses about which segments and transitions might differ in the amount of information they contain about an adjacent vowel. For instance, we mentioned above that liquids might contain more information than obstruents. When these differences are incorporated into the model, we can predict exactly the qualitative patterns of shortening observed in the production experiment.

Although based on known acoustic properties of segments, the hypotheses about recoverability that the formalism relies on largely consist of conjecture until we can confirm them empirically. Chapter 4 describes a perceptual experiment designed for this purpose. The experiment attempts to test hypotheses about the relative amount of 'vowel information' contained in various parts of the speech stream outside the vowel proper. In this study, subjects were asked to identify forward- and reverse-gated stimuli with truncated or removed vowels. The results display clear parallels to the compression asymmetries discovered in the production experiment. Subjects in general do significantly better at identifying adjacent vowels from liquids alone than from singleton obstruents alone. In onset position, where nasals induce incremental compression but obstruents do not, subjects show a significantly greater increase in sensitivity to vowel contrasts

as CV transitions are added back into the syllable for /nV/ sequences than they do for obstruent-vowel sequences. In coda position, where neither manner induces incremental compression, no such perceptual asymmetry exists.

The idea that patterns of duration and segmental overlap are governed in part by perceptual considerations is not new (Byrd 1994, Silverman 1995, Chitoran *et al.* 2002). In the formalization developed here, however, it becomes clear that the grammar of timing should have pervasive effects on phonotactic licensing cross-linguistically. In particular, many phonotactic generalizations can be explained with reference to *cue availability* and *perceptual distinctiveness* of contrasts (Steriade 1997, Flemming 2001, Wright 2004 *inter alia*). If temporal coordination affects and is affected by the same perceptual facts that drive phonotactics, we predict a wide range of duration-related effects on phonotactic licensing and repair strategies.

The available cues to any given contrast depend on language-particular patterns of phonetic realization (Steriade 1997, Gordon 2001, Jun 2002, Flemming 2008). In particular, cues to both the presence and the features of a stop in pre-stop position depend on how much it is overlapped with the following stop. If the two stops are very overlapped, the first one may not include an audible burst; if they are less overlapped, the burst will be audible; and if they are entirely non-overlapped, they will be separated by an open transition that provides cues to the presence and features of both stops. As such, the perceptibility of the first stop is largely a function of fine-grained timing relations between the two segments.

Chapter 5 extends the timing grammar from chapter 3 to include constraints on the number of contrasts in any given context and the distinctiveness of those contrasts. This extended grammar produces as output both categorical phonotactic patterns and fine-grained temporal representations. We show that it is capable of characterizing a host of facts involving consonant clusters and timing.

Like any cue-based approach, it predicts that a contrast will always be neutralized in environments where it is difficult to maintain the distinctiveness of that contrast before being neutralized in environments in which it is less difficult. The important sense of 'difficult' in the grammar here is 'produced with a relatively marked durational pattern'. From the general principles of timing developed in chapter 3 and the contrast constraints introduced here, we can derive the typology of stop-stop and stop-liquid clusters.

Given a detailed timing grammar, we can also analyze cases where the realization of contrasts varies language-internally, exemplified here by Tsou, Georgian, and Spanish. In Tsou and Georgian, the fine-grained temporal patterning of stop-stop clusters varies depending on context and place of articulation; these pattern are predicted by the timing grammar developed here. In Spanish, rhotics behave as sonorants in terms of phonotactic licensing, despite the fact that they are sometimes phonetically more similar to stops. The grammar developed here explains how the timing of these segments varies in order to preserve cues to a preceding obstruent; this contrasts with stops, which are not predicted to benefit from the same 'repair strategy'.

Finally, the grammar developed in chapter 5 makes strong typological predictions. One example of such a prediction: any language that licenses word-initial clusters with open transitions should also license coda consonants, but not *vice versa*. This is because the presence of consonants in the two positions is governed by a single constraint on contrasts, and a consonant cluster with open transitions will always produce a temporally more marked structure than a singleton coda consonant. On a first pass, this prediction appears to capture at least a strong tendency across languages. We examine seven languages that have been described as having CCV but no CVC syllables: Arabela, Cheke Holo, Lakhota, Mazateco, Pirahã, Piro, and Tsou. Two of these, Lakhota and Piro, are argued to include coda consonants; the other five do not license any clusters with open transitions.

A range of other cross-linguistic predictions that emerge from the particular theory of timing developed here are then elaborated and evaluated against the available empirical evidence. Most of these predictions involve some form of long-distance dependencies between phonological contrasts in different locations within a syllable. Some of the predictions are supported by a small number of attested patterns; others are unattested. In the final part of this chapter, we explore possible ways of constraining the formalism to eliminate unattested predictions.

Chapter 6 summarizes the preceding chapters and explores directions for future research.

2 An investigation of compression effects in English

2.1 Introduction

2.1.1 Overview

This chapter reports on an experiment that examines whether and how duration-trading relations manifest themselves in the English syllable. The general term used here to describe such relations is *compression effects*. The empirical and theoretical description of temporal coordination is of course significant in its own right; one goal of linguistics is to describe and analyze the world's languages, and timing relations at various levels of structure are part of the set of phenomena that must be described and analyzed. These phenomena are also of broad theoretical interest for several reasons. Asymmetries in compression across different contexts provide evidence about the division of labor between articulatory and auditory representations in language. And general properties of temporal coordination, be it articulatory or auditory, interact with the phonological licensing of contrast. As such, clarifying the empirical picture of compression effects will lead to greater understanding in other theoretical domains.

Complexity-driven compression effects have been approached from both an articulatory and an auditory standpoint; the two types of approach attribute compression effects to very different underlying principles, discussed in the next section. Given these differences in the underlying motivation for compression, the two approaches naturally make different predictions about where compression should be observed, and how much compression we should observe in different places. Testing these predictions will shed light on the nature of the representations that are relevant to timing and duration phenomena.

Timing phenomena are also of interest because they interact with systems of phonological contrast. For instance, one of the approaches to compression described here was developed to explain phonetic and phonological patterns of closed-syllable vowel shortening (Maddieson 1985, Flemming 2001).[1] In chapter 5, we highlight a variety of phenomena involving consonant clusters that both affect and are affected by compression. Given that these phenomena interact with temporal patterns, we can not hope to describe or analyze them fully without at least a partial theory of timing and duration. In the next section, we provide a sketch of two such theories.

2.1.2 Two approaches to compression

2.1.2.1 Articulatory approaches

One approach to compression treats it as essentially an epiphenomenon, one wholly determined by patterns of articulatory gestural coordination (Fowler 1981 *et seq.*, Browman & Goldstein 1990 *et seq.*, Nam *et al.* 2009). These theories include no mechanism for actively modulating the acoustic duration of a vowel, for instance. Rather, they include a small set of articulatory gestural coupling relations as primitives, and facts about acoustic duration emerge from those gestural relationships. Essentially, shortening happens when part of an articulatory gesture is encroached upon by an overlapping gesture.

The simplest version of this approach is laid out by Fowler (1983). The proposal is that consonant gestures are superimposed on the leading and trailing edges of vowel gestures.

---

[1] The terms *phonetic* and *phonological* refer here to non-neutralizing and contrast-neutralizing patterns of vowel shortening, respectively.

Essentially, the vowels form a 'scaffold' that can be used to support consonantal constrictions. This is shown in the figure below:

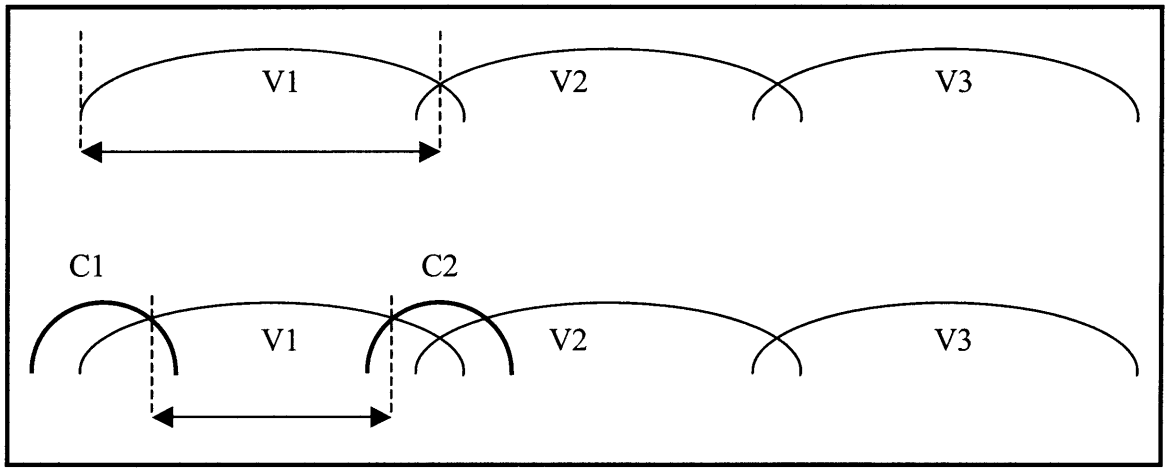

Figure 2.1. *A model featuring vowels as a gestural scaffold (top) and consonants as gestures overlaid on this scaffold (bottom). The introduction of consonantal gestures, the thicker arcs, has the effect of acoustically obscuring the part of the vowel gesture underneath those arcs. The arrows show this shortening. In this figure and those that follow, the vertical axis represents gestural activation.*

Assuming that the duration of vowel and consonant gestures remains constant between various contexts, this framework predicts pervasive compression effects. Every time a consonantal gesture is introduced into the speech stream, it masks part of a vowel gesture. The more consonants, the more masking. For instance, adding in C1 in figure 2.1 will tend to make the acoustic manifestation of V1 shorter; adding in C2 will tend to shorten the acoustic realizations of both V1 and V2; adding in a third consonant adjacent to C2 would result in even more shortening of V1 and V2.

This approach predicts that compression effects are more or less uniform across the grammar. Any time we add any kind of consonantal gesture in any position, it should drive vowel shortening. All segments should be alike in this regard, to the extent that they all at least partially mask the qualities of an adjacent vowel. Singleton consonants drive compression relative to no consonant; clusters drive compression relative to singletons.

This general theory could make different predictions about compression if it were coupled with a more specific theory of gestural alignment. Articulatory Phonology (henceforth AP) is just such a theory (Browman & Goldstein 1986 *et seq.*). One of the findings from this research program involves asymmetries in gestural alignment that depend on the number of consonantal gestures present and on the position of those gestures in the syllable.

Browman & Goldstein (1992) find that, for consonants in onset position, the beginning of the vowel gesture bears a constant temporal relationship to the temporal midpoint of the sequence of consonantal gestures (referred to as the *C-Center*). Across various singletons and clusters in onset position, what remains constant is not the temporal relationship between the beginnings of vowels and the beginnings of consonant complexes, but the relationship between the beginnings of vowels and C-centers. This is shown below.

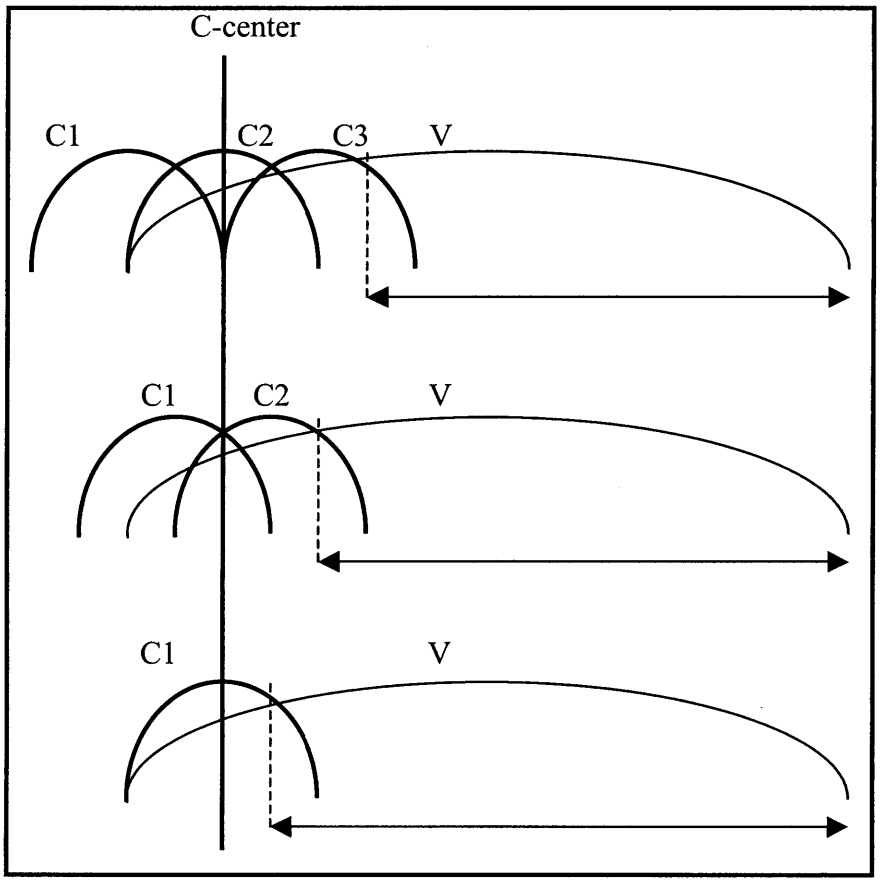

Figure 2.2. *The C-center effect. As more consonants are added into a syllable, the temporal relationship between the vowel onset and the C-center remains constant. The more consonant gestures are present, the more they impinge on the following vowel's gesture.*

It should be clear from figure 2.2 that the C-center effect will be accompanied by acoustic compression of the vowel. In order to keep the alignment of the C-center and vowel onset constant across clusters of increasing size, it is necessary for those clusters to impinge upon the following vowel gesture more as the number of consonant gestures in the cluster increases. Under the assumption that the duration of a vowel gesture remains fixed from one utterance to the next, and that acoustic vowel duration is roughly equal to duration of unmasked vowel gesture, this will result in acoustic shortening. This is illustrated by the arrows in figure 2.2.

In coda position, on the other hand, the c-center effect does not hold. In earlier versions of the AP model (e.g. Browman & Goldstein 1992), Browman & Goldstein report that the offset of the vowel gesture bears a constant temporal relationship to the beginning of the first consonant gesture in coda position, regardless of how many other consonant gestures might follow it. This is shown in figure 2.3.

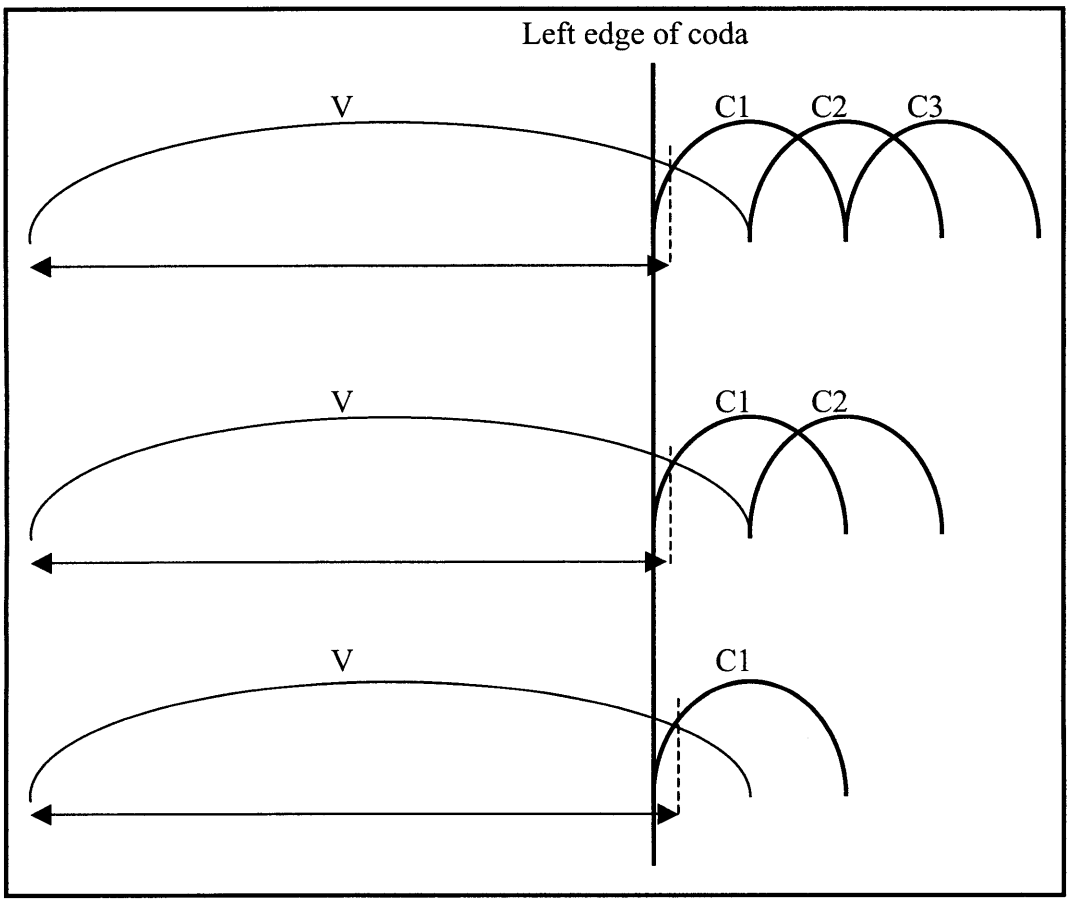

Figure 2.3. *Lack of a c-center effect in coda position. As more consonants are added to a syllable the relationship between vowel offset and onset of the first consonant gesture remains constant. Adding more consonant gestures will not impinge on the preceding vowel.*

The lack of a C-center effect in coda position means that there should be no additional vowel shortening as more and more consonantal gestures are added to a coda. Adding further consonantal gestures will not cause the preceding ones to impinge any more on the vowel gesture, and the acoustic duration of the vowel should not be affected.

This model differs from Fowler's in its predictions regarding compression. Fowler's model was formulated specifically with acoustic compression in mind, while this model was intended to account for an entirely different empirical phenomenon, that of the c-center effect. Despite the different motivation behind the AP model, it does still make predictions about acoustic compression: incremental with every added consonant in onset position, but constant in coda position. This prediction is made explicit in a later version of the AP model, which differs from the earlier implementation in the *explanation* of the c-center effect, but not its presence and absence by syllable position. According to Nam *et al.* (2009), 'adding Cs to a coda is predicted not to decrease the acoustic duration of the vowel'.

Broadly, then, these articulatory approaches predict that compression should obtain between syllables with different numbers of onset consonants, and possibly coda consonants as well. They predict a certain amount of uniformity in the phenomenon: compression arises whenever gestures overlap, regardless of the internal features of those gestures. In fact, this raises an important question of segmentation and coarticulation. Because the nature of masking and of coarticulation may vary between different linguistic sounds, the nature of what shortens may also vary. Presumably, when one sound more or less completely masks another (like a stop superimposed on a vowel), the audible duration of the masked sound should decrease. If the

masking relationship is only partial, producing an acoustic blend, than the portion of each sound that is not acoustically affected by the other should shorten. This suggests that it would be useful to examine compression effects on both the acoustic steady states of segments and the transitions between them.

2.1.2.2 Auditory approaches

An entirely different approach to compression effects emerges from investigations of *closed-syllable vowel shortening* (henceforth CSVS). In this phenomenon, which is widely attested cross-linguistically, vowels in closed syllables are observed to be shorter than vowels in open syllables. CSVS, then, is a specific sub-type of compression effect. The most frequent analysis of this pattern, whether explicit or implicit, is that it involves conflict between duration targets for smaller units such as segments and larger units such as moras, rimes, or syllables. The grammar, in this view, attempts to keep segments sufficiently long to be perceptible. It is also desirable to keep larger units sufficiently short to foster rapid and efficient communication, and to create at least a tendency toward evenly-spaced (or *isochronous*) sequences of these larger units. Long segments are good because they increase perceptibility, shorter syllables and greater overlap are good because they allow more contrasting units of information to be expressed over any given unit of time, and more isochronous syllables are good because they facilitate perception by inducing strong temporal expectations (Quené & Port 2005).

Analyses of CSVS often make reference to the idea that vowel compression is due to higher-level duration constraints, on a syllable or rime. Maddieson (1985), after arguing that closed-

syllable vowel shortening is widespread enough to be considered a near-universal, suggests that the phenomenon may itself be an argument for treating the syllable rime as a unit of timing. This implicitly suggests that compression effects are due to higher-level duration targets conflicting with lower-level ones. Myers (1987) invokes this trading approach in a phonological analysis of English closed-syllable vowel shortening. Flemming (2001) is a more recent and more explicit approach to closed-syllable vowel shortening in this vein. His model makes use of weighted constraints to characterize competing pressures on segment and syllable durations.

The general idea behind this approach can be captured with the metaphor of fitting small objects into a large container. As the number of small objects inside the container increases, the size of the objects and the size of the container come into conflict. We must either compress the small objects, or stretch the container, or both. This is illustrated below.

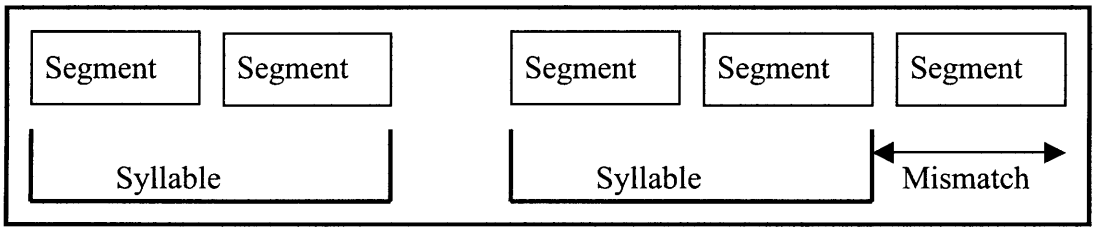

Figure 2.4. *Conflicting duration targets. As the number of segments inside a syllable increase, a mismatch arises between the target duration of the segments and the target duration of the syllable. The conflict can be resolved by shortening the syllables, lengthening the syllable, or both.*

In this approach, the auditory duration of lower-level and higher-level linguistic units is directly manipulated by the grammar. We use the syllable as the higher-level unit for this illustration and in much of what follows, but duration targets could in principle be associated with any of a number of larger units, such as the mora, rime, foot, or prosodic word.

The predictions of this general approach to compression are somewhat less constrained than the articulatory models outlined above. There are several reasons for this. First, the domain of compression depends upon which higher-level units have duration targets associated with them. If it is only the rime, for instance, then we expect compression driven by coda consonants but not onset consonants. If it is the syllable, on the other hand, we expect compression driven by all consonants.

Many researchers have written about CSVS because it is hypothesized to be of particular relevance to phonological phenomena involving vowel length contrasts. In many languages, CSVS is a gradient or phonetic phenomenon: closed syllables contain more or less the same vowels as open syllables, but shorter. In some languages, such as Egyptian Arabic (Broselow 1976) and Turkish (Clements & Keyser 1983), however, a contrast between long and short vowels in open syllables is neutralized to the short vowel in closed syllables.

Because these researchers have been concerned mainly with coda-driven vowel alternations, they tend not to consider the possibility of onset-driven CS. Correspondingly, we know of no attested language where vowel-length neutralization is conditioned by the presence of an onset consonant. If the explanation for vowel-length neutralization is truly to be found in compression

effects, then these theories should posit the rime (or possibly mora) as the higher-level duration target. This, in turn, would predict no onset-driven compression effects, and no interaction of onset consonants with vowel-length neutralization.

The second way in which the auditory approach might predict a greater variety of compression patterns than the articulatory approach concerns the nature of the representations that are involved in the phenomenon. If the grammar directly manipulates auditory duration, it may do so in a number of ways. For instance, if the motivation behind segmental duration targets is to maintain auditory perceptibility of those segments, then the grammar might introduce an absolute minimum duration threshold, which would constrain compression effects. The Klatt (1979) duration model proposes just such a parameter, although it is not explicitly concerned with compression.

A related prediction of the auditory approach is that compression effects might interact with other auditory properties of segments besides their duration. For instance, if sound α masks sound γ completely, but sound β masks γ only partially, the grammar may be sensitive to this distinction. Because duration targets in this approach are driven in part by auditory perceptibility, and overlap with sound β decreases the perceptibility of γ less than overlap with α does, we might predict less compression of γ adjacent to α than adjacent to β. Compare this to the articulatory approach, where auditory duration is an emergent side-effect of gestural overlap and is not part of phonetic representations. In that view, no actual shortening of segmental representations is taking place; only increased overlap. As such, we would not expect compression effects to be sensitive to the auditory properties of the segments involved.

The auditory approach, then, predicts that at least coda consonants will drive compression effects. It could also be stated in a way that predicts onset-driven compression, but this would create some problems for the phonological analysis of vowel-length neutralization. Furthermore, this approach predicts that compression effects might vary depending on the auditory properties of the segments involved.

2.1.3 Previous findings

We begin by introducing some terminology. Following Munhall *et al.* (1992), we refer to compression effects that are driven by increasing the number of segments in a string as *compensatory shortening* (henceforth CS). Because the current study examines compression in several contexts, it will be useful to introduce some terminology to describe those contexts. First, we distinguish between CS driven by the addition of segments to the onset of a syllable from CS driven by coda segments: *onset CS* vs. *coda CS*. We can also distinguish between CS observed in the comparison of syllables that contain one (consonantal) segment at the relevant periphery of the syllable (onset or coda) to syllables that contain no segments at the relevant periphery: *simplex CS*. For example, if we observe that the vowel is shorter in a CVC syllable than in a comparable CV syllable, it would be classified as simplex coda CS. Another case would be CS observed in the comparison of syllables that contain one (consonanal) segment at the periphery to syllables that contain more than one: *incremental CS*. For example, if we observe that the vowel is shorter in a CCVC syllable than a comparable CVC syllable, it would be classified as incremental onset CS. Many of the comparisons in this study examine incremental CS for pairs of items that involve the same consonant adjacent to a vowel, and differ in the presence or

absence of an additional consonant to the other side of the initial consonant. This includes pairs such as /nod/-/snod/ and /don/-/donz/. In cases where CS does obtain between such pairs, we say that the innermost consonant drives or induces incremental CS; this is something of a terminological shortcut or abbreviation. It reflects a hypothesis, to be made explicit later, that the innermost consonant is especially relevant to compression effects.

Several previous studies have found CS in various contexts in several languages. There are a few cases where different studies fail to agree. Here we summarize previous results that bear on the current discussion and explore ways to improve the methodology and analysis of previous studies. Fowler (1981) includes a brief review of literature on this subject before 1981.

Simplex coda CS, discussed here under the name closed-syllable vowel shortening, is widely attested cross-linguistically. These investigations tend to be concerned with rime-driven phenomena such as syllable weight (Broselow *et al.* 1997) and contour-tone licensing (Zhang 2004); as such they generally don't touch on any other type of compression effect. Maddieson (1985) gives an extensive review of languages where the phenomenon has been attested; he tentatively proposes that it is a universal tendency.

A number of studies find simplex and incremental CS in both onset and coda position. Lindblom & Rapp (1973) show this for Swedish; they report that the coda effect is stronger. Fowler (1983) reports this pattern for English. Clements & Hertz (1996) report that English displays simplex CS for segments in what they call 'the extended nucleus' of a syllable, which includes voiced transitions preceding and following the vowel, following glides, and following liquids. For

instance, they find that the steady-state vowel is much shorter in *bait* than *wait*, presumably because the transition is so much longer in *wait*; similarly, the /a/ portion of *tide* is shorter than that of *Todd*, because the nucleus is more 'crowded' with the diphthongal offglide present. Munhall *et al.* (1992) find incremental coda CS for obstruent-obstruent clusters. The effects are generally rather small (the largest is 36 ms but most comparisons are on the order of 3-10 ms) and vary between subjects.

Although none of these studies examines differences in CS across consonant manners, there are a few studies that report some relevant data in this regard.

Van Santen (1992) reports on a large corpus study of English. He finds differences in preceding vowel duration depending on the following consonant; for instance, voicing and frication correlate with longer preceding vowels; /r/ is preceded by extremely short vowels. He also finds a small but significant incremental onset CS effect (about 10 ms) for obstruent-liquid clusters. Results are not reported for obstruent-obstruent clusters, and not enough data was available to assess obstruent-nasal clusters. The study finds significant simplex CS for onsets and codas, but there is not enough data to distinguish between classes of consonant in this regard.

Waals (1999) examines the durational properties of various consonants and clusters in Dutch, including some data that bear on the question of CS. She finds that, in onset position, consonants in clusters are shortened relative to singleton counterparts. Compression disproportionately affects higher-sonority segments relative to lower-sonority ones. In coda position, she finds incremental vowel shortening between singleton and cluster codas for all segment types, and

possibly for two vs. three coda consonants. This effect is much larger for vowels preceding /r/ than those preceding other consonants. For long vowels, the effect is largest for those preceding liquids, intermediate with /n/, and smallest with obstruents.

Katz (2008) finds simplex CS in both onset and coda position in English. The study finds incremental CS for /l/ in both onset and coda position, but not for obstruents in either position. There is an incremental CS effect for /n/ in onset position, but it varies between subjects and is only marginally significant. Nasals in coda position were not examined.

One series of studies fails to find convincing evidence for compression effects in English. Crystal & House (1982, 1988, 1990) report duration measurements from a study of 6 speakers reading a short script. They find no strong evidence for compression effects in stressed syllables, but some evidence in unstressed syllables. They report that the sonorant/obstruent distinction has no effect on the duration of a preceding vowel.

Taken as a whole, the literature suggests that CS is present in some form in various contexts, but many questions still remain, and some studies have failed to find any effect at all. There is some evidence that CS may differ across manners of consonant. Only two of the studies described above directly compare onset and coda CS; some of the studies test only simplex or only incremental CS. Most of the studies make no attempt to compare different manners of consonant, and the ones that do often don't cross these differences with syllable position or number of consonants.

The current experiment was designed to test for CS across a range of consonants and contexts in English. The study also attempts to avoid some of the methodological shortcomings of previous studies. These shortcomings include a small subject pool, lack of appropriate (or any) statistical analysis, failure to distinguish between types of vowel and consonant, elicitation of artificially rhythmic speech using a single carrier phrase or a metronome, and a lack of clarity or precision in characterizing segment boundaries. It is not the case that all of the studies described above suffer from all of these problems, but each of those studies suffers from at least one of them.

The current study reports results from six speakers. While this is not a particularly large number, it is more than any of the studies reported here except for Crystal & House (1982 *et seq.*), which pooled counts across subjects. The current study involved a large number of measurements taken by hand; as such, the time required to analyze materials grows hugely with each additional subject. As we'll see, six speakers are enough to get significant and coherent results.

The data were analyzed with linear mixed effects regression models, which are described in the next section. These models allow us to ask questions about fine-grained differences in duration in a principled, quantitative way, and to assess the reliability of the answers we find.

The materials include three vowels and four 'series' of consonant, meaning that the same consonant is elicited as a singleton and in a cluster. The four series target two liquids, a nasal, and two obstruents (the obstruent series could not be completely identical across onset and coda position). The statistical analysis starts with the assumption that each consonant series (and each vowel) may have a different CS pattern and proceeds by eliminating model parameters to

generalize across segments. In this way, we avoid making unwarranted assumptions about which segments are equivalent with respect to CS; prior literature offers some evidence that different manners of consonant, at least, may differ in this regard.

The materials were elicited with a set of different carrier sentences, which were broken up by prosodically, syntactically, and semantically diverse filler sentences. This results in speech that is less rhythmically constrained than repeating one phrase over and over again. The drawback is that the variance in the study is larger than in a more constrained task, which could obscure small effects. Because there are already substantial findings about CS in repetitive and isochronous speech, however, it is now desirable to see whether the findings extend to more naturalistic speech. The next section also reports several methodological and statistical attempts to control for speech rate and prosodic phrasing.

The problem of boundary criteria for segments is a more vexing one. It is clearly difficult to find a set of objective criteria for drawing a boundary between vowels and liquids, for instance. The same uncertainty also arises in cases that would seem to be relatively clear, such as vowel-obstruent boundaries. For instance, it seems obvious that to draw the boundary between a vowel and /s/ we would look for the dividing line between periodic voicing on one side and aperiodic frication on the other. In reality, such a dividing line is often not present in running speech: high-frequency noise creeps into periodic phonation and the one gradually subsides into the other (and in the case where the vowel precedes the fricative, the transition often goes through a phase of breathy or voiceless /h/-like noise in between the two segments). Because segments are coarticulated, there is almost never in principle a clear point in the signal where one segment

ends and the next begins. Even in relatively clear-cut cases such as stop-vowel boundaries, the signal changes from acoustic properties characteristic of one segment to those of another over a non-negligible period of time.

The solution adopted here is in the spirit of the *phone-and-transition* model advocated by Hertz (1991) though it differs in some details. In a sequence of two segments, the acoustic signal is segmented into the steady state of the first segment, the transition between the two segments, and the steady state of the second segment. Each boundary is selected using a particular acoustic landmark or combination of landmarks. For instance, in a token of /lɑ/, the transition begins when F1, which remains relatively steady internal to the /l/, begins to rise; the end of the transition and the onset of the vowel steady state is marked when F1 stops rising in the vowel. Even this model is an idealization; there is often no clear single point in the acoustic signal where a formant or other acoustic entity goes from some slope to no slope. The experimenter attempted to identify a small portion of the signal as containing the boundary; within that portion, exact boundary selection was often guided by the Praat (Boersma & Weenink) formant tracker. There is undoubtedly measurement error in the data; the hope is that it is essentially random and should not unduly affect the duration of some segments more than others.

The main objective for the segmentation strategy is to delimit intervals that are comparable across items that differ in the number of target consonants. Although the boundary between transition and vowel proper in the /lɑ/ case discussed above may not correspond to any psychologically real boundary between two symbols, it at least gives us an acoustic landmark that can be compared to the same landmark in tokens of /glɑ/ (as well as /ɑl/, and /nɑ/). If we find

that the interval of vowel with steady F1 in /glɑ/ is shorter than that in /lɑ/, it entails that there is incremental onset CS for /l/. On the other hand, this boundary won't be strictly comparable to the one in /rɑ/, where the comparable boundary tracks F3 rather than F1. If we find that the intervals delimited by such boundaries differ in duration, the most we can say is that the interval of vowel with steady F3 in /rɑ/ is shorter than the interval of vowel with steady F1 in /lɑ/. Similarly, the marked boundaries in /lɑ/ are not strictly comparable to those in /lɪ/, which track F2, or in /p^hɑ/, which track aspiration.

Note that the term 'steady-state' is used here as a label for an interval marked in a particular way; this does not imply that all spectral properties are static within the interval. For instance, the 'steady-state' of an /o/ may be segmented on the basis of F2 movement; within the marked steady-state, there may be a fair bit of F1 movement. What 'steady-state' really means is, in this case, something like 'the interval beginning/ending at a local F2 plateau and extending to the fixed /d/ on the other side of the vowel'.

This method results in boundaries that may not correspond to what we intuitively think of as 'the' boundary between two segments. For instance, in the /lɑ/ token, an approach that tries to mark the true boundary between /l/ and /ɑ/ would likely place it somewhere inside the segment marked as a transition in the current study. As such, some of the vowel and consonant durations reported here may differ from previous studies or from accepted facts about English vowel duration (e.g. the period marked as vowel proper is far shorter before liquids than before

voiceless obstruents). One of the points this study should reiterate and drive home is that the notion of a boundary between segments is not particularly well-founded. The ways in which segments overlap are an important part of duration patterning, and need to be measured and analyzed in any work that purports to describe these patterns.

Besides an explicit, objective, and replicable set of boundary criteria, the segmentation method used here offers several analytical advantages. Previous studies on CS have generally marked off boundaries, equated them with segments, and shown that some segment shortens from one item to another (Clements & Hertz 1996 and the articulatory data from Munhall *et al.* 1992 are exceptions). The current study will allow us to see in greater detail exactly *what* shortens in CS; no theory of temporal coordination is complete until it has addressed this issue.

2.2 Methods

2.2.1 Materials

The 'target' materials consisted of every phonotactically legal combination of the vowels {ɪ, ɑ, o} with: the consonants{r, l, n, ∅} in onset and coda position; /p/ in onset position; /s/ in coda position; the clusters {br, gl, sn, sp} in onset position; and the clusters {rb, lb, nz, sp} in coda position. Each item contained a 'fixed' consonant /d/ at the opposite edge of the syllable/word from the one being manipulated. The number of logically possible combinations is 54. Three of these (/dɪ/, /dɪr/, and /dɪrb/) are phonotactically illegal in English. /dosp/ is arguably ill-formed as well, because there are no English words with a tense vowel followed by a cluster of obstruents

where one is non-coronal: *toast* and *cusp* are OK, but **toasp* may not be. Because this case is not as clear-cut as the obviously ill-formed words mentioned above, it was included. This left a total of 51 target syllables/words. Of these, 24 correspond to existing English words (if the slang word *diss/dis* is counted); the remaining 27 are nonce-words.

IPA	Orth.	IPA	Orth.	IPA	Orth.
ɑd	od	od	oad	dɪn	dinn
lɑd	lod	lod	lode	don	doan
glɑd	glod	glod	gload	dɑnz	donz
dɑ	dah	do	doh	dɪnz	dinz
dɑl	dall	dol	dole	donz	doanze
dɑlb	dalb	dolb	dolb	pɑd	podd
rɑd	rodd	rod	rode	pɪd	pid
brɑd	brod	brod	brode	pod	poad
dɑr	dar	dor	dore	spɑd	spod
dɑrb	darb	dorb	dorb	spɪd	spid
ɪd	idd	nɑd	nodd	spod	spoad
lɪd	lidd	nɪd	nid	dɑs	doss
glɪd	glid	nod	noad	dɪs	diss
dɪl	dil	snɑd	snod	dos	doase
dɪlb	dilb	snɪd	snid	dɑsp	dosp
rɪd	ridd	snod	snoad	dɪsp	disp
brɪd	brid	dɑn	don	dosp	doasp

Table 2.1. *Phonetic and orthographic representations of the 51 target words elicited in the experiment.*

These items were chosen to include a variety of consonant manners, to compare singletons and clusters, and to compare onsets and codas. It was impossible to satisfy all of these goals perfectly. English only realizes voiceless singleton stop onsets in stressed syllables as aspirated, but their counterparts in /sp/ clusters are unaspirated. As such, this is not a minimal pair (because

the members differ in both aspiration and the presence of /s/). The only possible cluster in onset position with /n/ as the second consonant is /sn/; in coda position, however, we tested /nz/ instead of /ns/. This is because voiceless obstruents induce radical shortening of a preceding vowel (Peterson & Lehiste 1960). Any vowel shortening we uncovered in an /ns/ sequence couldn't be attributed with certainty to CS; it might also be a property of the voicing contrast.

Wherever possible, items were assigned orthographic respresentations that do not correspond to English words. The only exceptions are *rode*, *lode*, *don*, and possibly *diss* (meaning 'disrespect') and *doh* (an exclamation of dismay associated with Homer Simpson). Some of the words unavoidably were assigned unusual or ambiguous orthographic representations. The pronunciation of nine such words was demonstrated to subjects at the beginning of the experiment.

In addition to the target items, 39 filler words were included in the reading session. These were also monosyllables, with different consonants and vowels than the target items, including some consonant clusters. *Freave*, *skay*, and *jeg* are examples of filler words used in the experiment.

The experiment included 17 target carrier sentences and 13 filler carrier sentences. The target carrier sentences were strictly controlled for prosodic factors. Each sentence was nine syllables long, of the form [[*X*] [*Y the Z W*]], where: *X* is a trochaic first name; *Y* is a past tense monosyllabic verb; *Z* is the target item; and *W* is a four-syllable modifier, beginning with a preposition and containing one noun (with one exception, mentioned later). *Thomas bought the*

dore at a yard sale and *Dustin got the snid off of E-bay* are examples of target sentences used in the study.

The expectation was that, given their identical syllable count and syntactic structure, the target sentences would elicit comparable prosodic structures across utterances. The target word itself is determiner- or noun-phrase-final but not utterance-final in these sentences, and is expected to be produced with a pitch accent.

Because the target sentences are so similar rhythmically, filler sentences were formulated to disrupt the sense of repetition, which could result in effects of isochrony or parallelism not characteristic of natural speech. The filler sentences vary in their length, syntactic structure, and illocutionary force. They include questions, statements of opinion, and direct and indirect commands. *The yeam is poisonous, right?* and *This jutch wouldn't be a bad thing to buy* are examples of filler sentences used in the study.

There were 90 total experimental items (target and filler) to pair with 30 carrier sentences; each experimental block of 30 trials included one third of the experimental items and each carrier sentence. Pairings of experimental item and carrier sentence were randomized, as was the order of trials inside each block of 30. The randomized sentences were presented to subjects on a computer screen. They were asked to 'read each sentence in as natural a manner as possible' before pressing a button to move to the next sentence. They were given the opportunity to take breaks after each block of 90. There were 4 repetitions of each experimental item (paired with a

randomly selected carrier sentence each time) for a total of 360 utterances. The experiment ran between 30 and 45 minutes for all subjects.

2.2.2 Subjects

Subjects were 6 native speakers of American English, 4 female, 2 male, all between 21 and 31 years old. None reported being diagnosed with any speech, reading, or hearing disorders. Three were from Massachusetts; the other three from New York, North Carolina, and Minnesota. Subjects were debriefed after their recording sessions; none reported knowing what the experiment was 'about'.

2.2.3 Recording

Subjects were recorded in a sound-attenuated booth inside the MIT phonetics laboratory. They were outfitted with a head-mounted condenser microphone, placed at an oblique angle to the lips. They read the experimental sentences off of a computer screen. The utterances were recorded in mono at 44.1 kHz with the Amadeus software and saved to .aiff files.

2.2.4 Measurement

The recordings were cut into smaller files and annotated for duration by hand using the Praat software (Boersma & Weenink). In the descriptions that follow, I make reference to 'acoustic values' as a catch-all term for the variety of acoustic properties used in segmentation. Full details

of what these properties are and how they were used can be found in Appendix 2A. For all words, the following regions were marked:

- Vowel proper: the portion of vowel from the innermost edge of the fixed consonant to the first point where acoustic values begin to slope noticeably toward characteristic values for the target consonant
- Fixed consonant: for onset /d/, the region extending from an abrupt drop in (or cessation of) energy in the preceding schwa to the onset of periodic voicing in the vowel of the target word; for coda /d/, the region extending from an abrupt drop in (or cessation of) energy to just after the following release burst. In cases where the /d/ was realized as a tap, the offset was marked after the abrupt drop in energy and formants around the tap.

All target words appeared in between two vowels: they were preceded by the vowel in *the*, most often realized as a schwa; and followed by the initial vowel of a preposition, which varied across carrier sentences. The terms *C1* and *C2* will be used to refer to the innermost and outermost target consonant, respectively. So, for instance, /dolb/ has /l/ as C1 and /b/ as C2; /brod/ has /r/ as C1 and /b/ as C2.

For words with target consonants (not VC or CV words), the following regions were also marked where applicable:

- Transition: the region extending from the vowel proper to the steady-state portion of the adjacent target consonant.

- C1: the region extending from the onset or offset of the innermost (i.e., adjacent to the vowel) target consonant to the first point where acoustic values begin to slope noticeably toward characteristic values for the vowel.
- C2: The region extending from the onset or offset of C1 to the onset or offset of the outermost (i.e., not adjacent to the vowel) consonant.

No attempt was made to place a boundary between the vowel proper and the transition to the fixed consonant /d/; rather, the vowel proper measurement incorporates this transition (though not the closure or release of /d/). Because all of the items have fixed /d/ in them, variation between tokens with regard to this transition should affect all tokens equally, on average, and should show up in the results only as random noise.

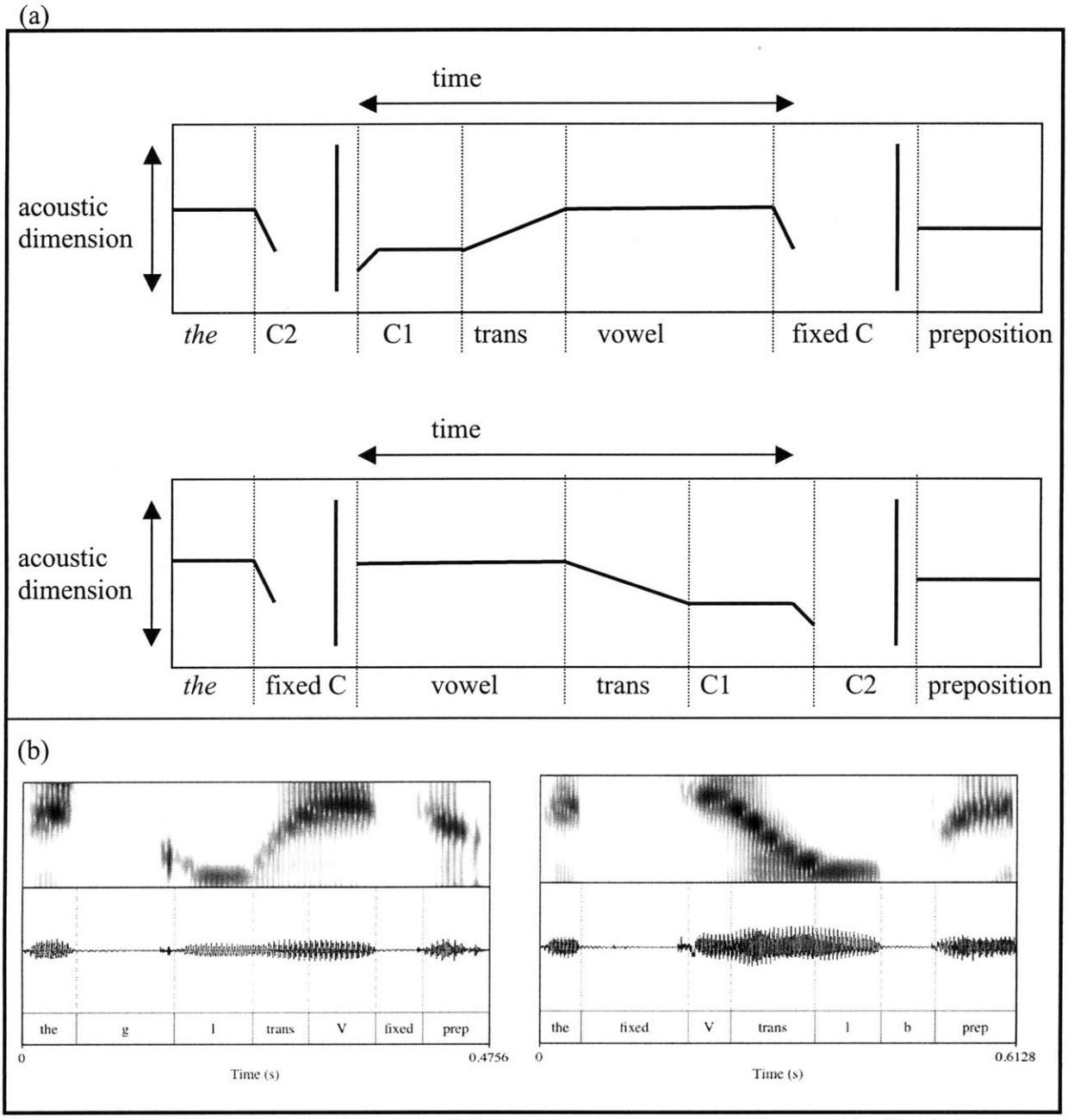

Figure 2.5(a). *Schematic illustration of duration measurements, using an idealized single acoustic dimension that tracks all boundaries. In reality, the relevant acoustic dimension would be different for different boundaries. The top illustration shows a CCVd word; the bottom a dVCC word.* **(b).** *Actual implementation of segmentation strategy in a token of /glɪd/ (left) and /dɪlb/ (right) from subject PC; spectrogram shows F2.*

The strength of the prosodic boundary following the target word varied somewhat, both between and within speakers. Realizations ranged from no noticeable temporal discontinuity to a full pause, sometimes including a schwa-like excrescence following the final consonant. Every recorded token included what could be considered a pitch accent on the target word; the most common realization would be labelled as a H* tone followed by a L- phrasal tone in the TOBI model (Silverman *et al.* 1992). Because there were few or no instances of unaccented or deaccented target words, no attempt was made to systematically transcribe the prosody of the materials. One target sentence, which ended with the modifier *all in one batch*, was consistently produced with a slightly different prosodic structure than the other sentences. This sentence tended to have a pitch accent on *all*, whereas the other sentences contained in the corresponding position an unaccented preposition. In addition, almost all tokens of this sentence included a noticeable temporal discontinuity or pause between the target word and *all*. We return to this irregularity in the results section.

2.2.5 Analysis

Separate models were constructed for each of three dependent variables: duration of the steady-state vowel, duration of the CV/VC transition, and duration of steady-state C1. The data were analyzed with linear mixed effects regression models.[2] This type of model offers several advantages over the repeated measures ANOVA models that are common in speech and language research (Quené & van den Bergh 2004, Baayen *et al.* 2008).

[2] The discussion of mixed-effects models draws heavily on Quené & van den Bergh 2004; mathematical concepts that are only mentioned here are explained more fully in that paper.

A repeated measures model makes assumptions about the distribution of data points; it assumes that variance is comparable between conditions (*homogeneity of variance*) and that, roughly speaking, co-variance between each pair of conditions is comparable to every other pair (*sphericity*). Mixed effects models don't rely on either assumption, as the variance-covariance matrix is modeled directly from the data, rather than being taken as a given.

A repeated measures model can't accomodate missing data points. Essentially, if we fail to obtain any one observation from a subject, we must discard all of that subject's data or find a way to impute those data. Mixed effects models are robust to missing data points.

A repeated measures model can only accommodate one random effect at a time. In the case of nested or crossed random variables, there are two possibilities: the experiment must be counter-balanced; or the analysis must incorporate a separate model for each random variable, with an approximate criterion for determining statistical significance based on test statistics from each of the models (Raaijmakers *et al.* 1999). Mixed effects models can in principle incorporate many random effects at once.

Finally, mixed effects models appear to be more powerful than comparable repeated measures models. This means that they are better suited to detecting and quantifying meaningful trends in a data set. This is probably the most important difference between the two types of model.
The models reported here are the end product of a hierarchical backward elimination procedure. The process begins with a baseline model that includes a separate parameter for each item in the experiment. Such a model corresponds to a theory of temporal coordination where each lexical

item (or perhaps each bigram) is stored in memory with its own idiosyncratic timing pattern, and there are not necessarily any useful generalizations to be made about similarities in timing between words with similar segments. This is an extremely weak hypothesis, in the sense that it makes fewer predictions than a theory which includes equivalence classes (segments, features, cues) internal to lexical items. Successively stronger theories are then tested by removing parameters or blocks of parameters from the model. This corresponds to modifying our hypothesis to include ever more general equivalence classes. Checking how much these removals decrease the fit of the model to the data will tell us how much empirical coverage we lose by strengthening the hypothesis.

This hierarchical process was adopted to allow examination of many possible influences on CS without wildly inflating the number of *ad hoc* parameters in the model and overfitting the data. As summarized in section 1, a lot remains to be learned about where CS occurs and how it differs according to context. Fortunately, we have a theory of linguistic units to guide our research that includes at least vowels, consonants, manner features, and linear order. Hierarchical modeling allows us to ask about contrasts that are based on linguistic units, rather than haphazardly searching for predictors that improve model fit. This avoids many of the pitfalls associated with those stepwise or sequential regression procedures where predictors are selected *post hoc* on the sole basis of their quantitative properties.

The baseline model includes *random effects* of subject identity and carrier sentence. These are variables whose levels are sampled from the larger population at random, without covering every possible level in those populations (e.g. the 17 target sentences used in this experiment are just a

tiny fraction of all English sentences sharing a particular set of syntactic and prosodic properties).

The model included *fixed effects* of two kinds: level-defining effects that were manipulated to create the different experimental conditions; and normalizing effects that attempt to control for differences in speech rate, prosodic structure, allophony, and any other phenomena that might differ between utterances. Note that this distinction into two types of fixed effect is purely for expository purposes; the variables are treated exactly the same by the model.

The normalizing effects pertain to several different aspects of the materials. Lexical status ({word, non-word}) and frequency (natural logarithm of values from the CELEX database)[3] pertain to the familiarity of each item. Trial (how far along in the experiment the item was uttered) pertains to possible changes in speech rate, familiarity, and concentration as the experiment progresses. For items in the onset condition, the allophonic status of word-final fixed /d/ ({flapped, non-flapped}) pertains to speech rate and prosodic phrasing,[4] as does the duration of the fixed consonant in both onset and coda positions. Two variables corresponding to allophonic properties of VC words will be discussed in section 2.3.

---

[3] Two English words included in this study, *rid* and *diss*, have no listing in CELEX. They were assigned the mean log frequency of the other existing English words in the experiment. This solution was adopted because we don't believe that these words are vanishingly rare. We suspect that the omission of *rid* is some type of an editing or compilation error, and that *dis/diss* is either too recent a coinage or too rare in written language to appear in CELEX.

[4] Note that this variable was coded 0 if there was a visible or audible burst in the realization of /d/, 1 otherwise; this may not correspond exactly to intuitions about what is and is not a flap, but it is at least a concrete and replicable criterion.

The level-defining effects are vowel ({ɪ, ɑ, o}), C1 quality ({rhotic, lateral, nasal, obstruent}), syllabic position of the target consonant(s) ({onset, coda}), and number of target consonants ({0, 1, 2}). The baseline model, then, would be one that includes all 4-way interactions between these variables (with the exception of the impossible items mentioned above). Removing higher-order interactions from the model generalizes across classes of item, creating a stronger hypothesis. The statistical significance of the higher-order interactions amounts to a metric of how much we've damaged the empirical coverage of our hypothesis by making it more general.

At each step, the significance of the fixed effects was assessed using Markov chain Monte Carlo (MCMC) sampling. Roughly speaking, this procedure generates hypothetical sets of parameters over and over again, then compares these parameters to the actual ones the model has fitted to the data, in order to assess the probability of obtaining such extreme parameters by chance. Baayen *et al.* (2008) give a more detailed description of this procedure.

Non-significant fixed effects were removed level by level if they included a term for number of target consonants. These are the parameters that test whether CS is present, and whether it varies from one context to another. All fixed effects were retained if they did not include a number-of-consonants term. Because 1 consonant (CVC) was used as the reference level, these parameters define baselines (generally durations in a CVC syllable) for each condition, against which the CS parameters of the model are tested. Hence, even if they are non-significant, retaining them can only increase the accuracy of the estimated CS effects. After each elimination step, MCMC simulation was repeated for the reduced model.

The significance of random effects is calculated differently in a mixed-effects model. To check for subject interactions with a fixed effect, for example, we must include a *by-subject random slope* for the fixed effect of interest. We then check how much this parameter improves the model fit by comparing the performance of the reduced and expanded model using a chi-squared test of the likelihood ratio. After the fixed effects in the model had been reduced by the procedure described above, by-subject random slopes were tested. All significant fixed main effects were examined; if the by-subject slope for two main effects both resulted in significant improvement of the model, the by-subject slope of their interaction was also tested.

In some cases, including subject interactions changed the estimated values of fixed effects. As there is currently no way of running an MCMC simulation on a model that includes by-subject random slopes, the significance of these changes could not be assessed with certainty. There is, however, an approximate way of gauging whether effects are likely to be significant without running an MCMC simulation.[5] This is to check the value of the *t* statistic for each parameter. In an experiment with few observations and few degrees of freedom, this statistic is anti-conservative (it inflates the probability of Type I error, rejecting a true null hypothesis). At relatively great degrees of freedom, however, the *t*-statistic converges on the standard normal distribution. As such, we can roughly gauge whether an effect is significant at the $\alpha = 0.05$ level by checking whether the absolute value of the *t*-statistic is greater than 2. There is currently no generally agreed-upon method for determining the degrees of freedom for a mixed effects model. The upper bound of estimated degrees of freedom for the current experiment (equal to the number of observations minus the number of fixed effect parameters) is greater than 1,100. None

[5] This argument is from Baayen *et al.* (2008).

of the changes after random slopes were added resulted in any fixed parameter moving from a t-value greater than 2 to one less than 2, or *vice versa*, so we can tentatively conclude that the addition of subject effects doesn't qualitatively change the nature of the results.

In what follows, all fixed effects will be reported with an effect size and p-value from MCMC sampling. All random effects will be reported with an effect size, chi-squared statistic, and p-value from a test of likelihood ratios. In cases where MCMC sampling is not available, fixed effects will be reported with an effect size and *t* statistic.

Before statistical analysis, the data were centered around 0 and normalized using a *z* transformation for each subject. This transformation characterizes data points by how many standard deviations they lie above (positive values) or below (negative values) the mean. Effect sizes, then, are in standard deviations; in the text, they are translated back into a range of ms values for ease of comprehension. These ms values represent the range obtained by multiplying the z-transformed effect size by the smallest and largest subject standard deviations.

2.3 Results

2.3.1 Simplex CS

2.3.1.1 VC syllables

The VC syllables in the experiment were realized with substantial variation pertaining to the presence and nature of a glottal constriction at the beginning of the item. Some tokens included a realization of *the* as /ði/, with a modally-voiced transition between /i/ and the target vowel; other

tokens included full glottal closure following a schwa in *the*, with near-immediate modal voicing of the target vowel upon release. The majority of tokens fell on a continuum between these two endpoints. For instance, some tokens included a creaky-voiced transition between the two vowels. In some cases, this was preceded or punctuated by fairly long closures; in some cases glottal pulses were irregular but more or less continuous. Illustrative examples of various realizations are shown in figure 2.6.

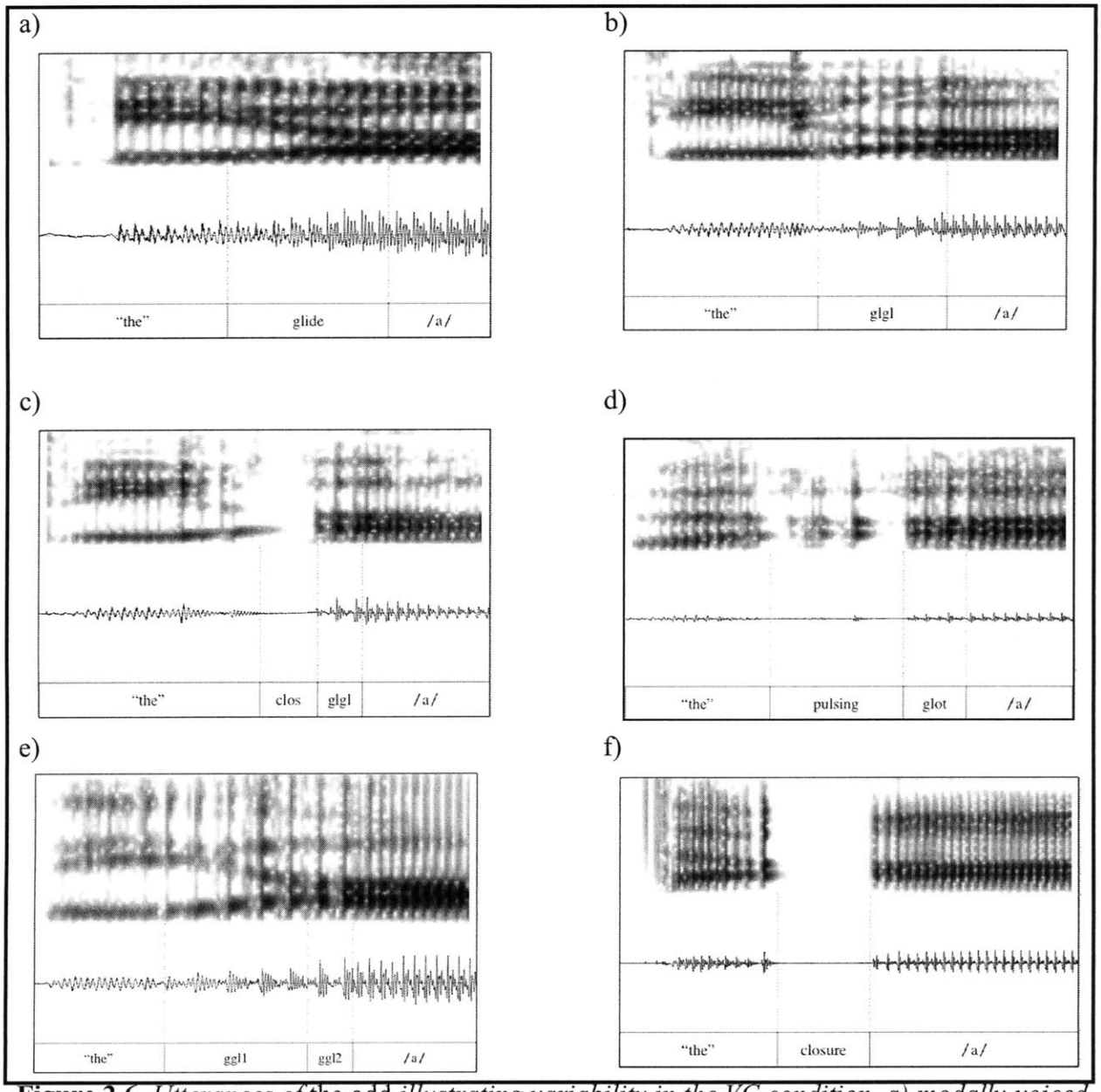

Figure 2.6. *Utterances of* the odd *illustrating variability in the VC condition. a) modally-voiced transition. b) creaky-voiced transition. c) closure followed by creaky onset. d) intermittent glottal pulsing followed by creaky onset. e) creaky transition and creaky steady state followed by modal voicing. f) full glottal stop.*

This variability raises the question of what should 'count' as vowel duration in these items, both psychologically and for analytical purposes. Investigating various metrics of duration is desirable

both for what it reveals about the temporal organization of these materials and for discovering the most consistent and principled metric to use in comparisons with other materials. Three different metrics were investigated:

- **m1**: only the portion of the vowel with modal voicing and steady formants
- **m2**: the portion of the vowel with steady formants, regardless of glottalization
- **m3**: the entire portion of the vowel with visible formant structure

The third metric produced more consistent results than the other two. As a preliminary, the standard deviation was computed for each metric within each vowel; m3 produced the smallest values, indicating less variability. This is despite the fact that the absolute numbers for m3 are the largest of the three metrics.

	a	i	o
m1	58.2	45.1	39.2
m2	57.4	42.5	43.9
m3	34.3	36.4	31.2

Table 2.2. *Standard deviations, in ms, for each metric and each vowel.*

As a purely practical matter, m3 was adopted as the measure for VC items in all further statistical modeling. The smaller variability under this metric will make it easier to compare these items to others in the experiment. The nature of the relationship between duration and onset quality in these items is also of theoretical interest, however. Comparison just within this class of item reveals something about the nature of CS.

To examine patterns in greater detail, VC materials were split into five classifications of onset quality, corresponding to the tokens in figure 2.6 (a-c, e-f); the realization shown in 2.6 (d) was

not observed often enough to conclude anything about it. The classes will be referred to as *glide* (2.2a), *creaky transition* (2.2b), *creaky steady state* (2.2c), *2-part creak* (2.2e), and *stop* (2.2f). Figure 2.7 shows the average duration of each part of the VC syllables, separated into classes.

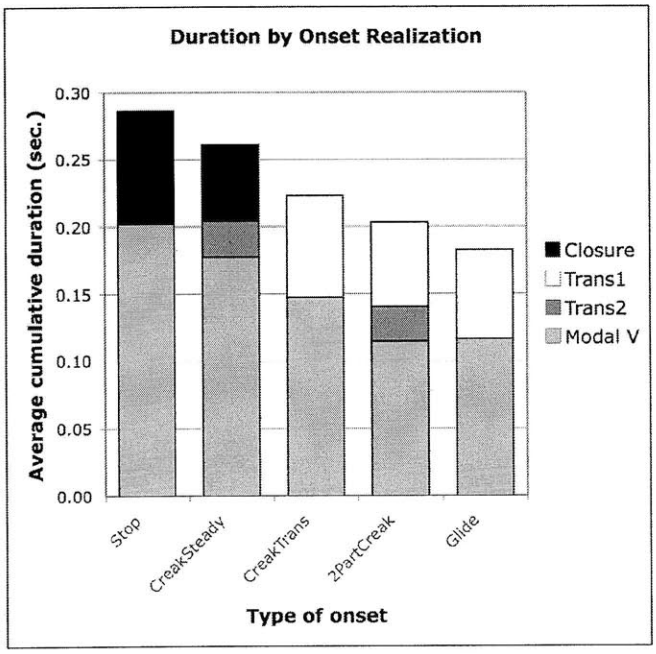

Figure 2.7. *Duration, portion-by-portion, of VC syllables, separated into onset classes. 'Trans2' refers to creaky steady state; 'Trans1' refers to a transition with moving formants.*

The most obvious pattern in figure 2.7 is that modal, steady-state vowel duration is longer in the classes where it is not preceded by formant transitions (or, alternatively, in the classes where it *is* preceded by closure). This suggests that the transitions 'count' at least partially as vowel duration in whatever sense is relevant to a speaker's temporal coordination. Correspondingly, the metric which includes these transitions, m3, results in a more uniform characterization of vowel duration in VC syllables than the other 2 metrics. This is shown in figure 2.8.

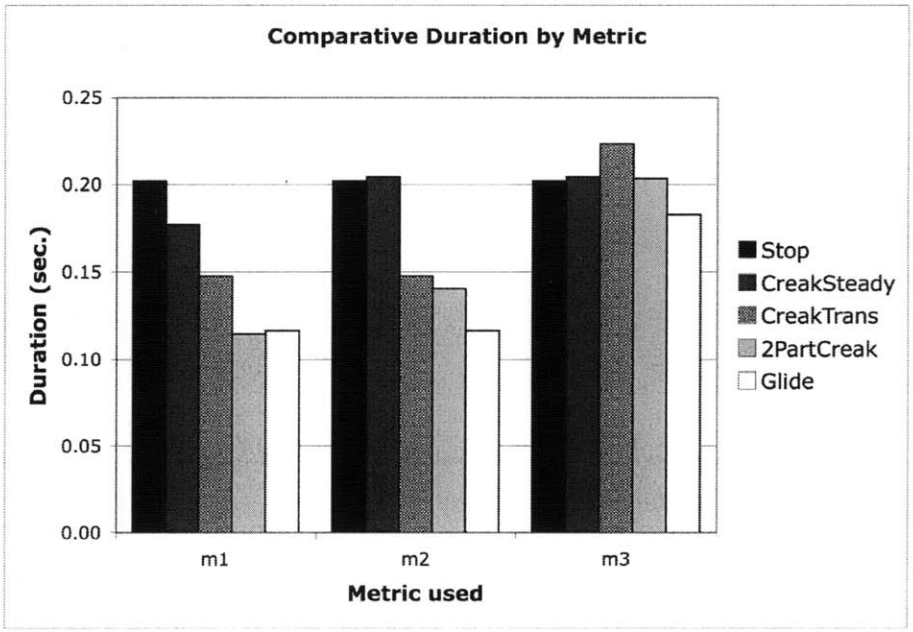

Figure 2.8. *Vowel duration across onset class and metric.*

Note that, even on metric m3, duration is not perfectly uniform across classes. This metric characterizes the creaky transition realizations as having somewhat longer vowels than the others, and the glide realizations as having somewhat shorter vowels than the others. One hypothesis would be that transitions don't count entirely as a part of the vowel, but that different types of transitions count in different proportions. This property will figure prominently in the analysis of other phenomena in the experiment. Metric m3 was used for the statistical model reported here, with separate variables corresponding to the presence of creaky transition and glide realizations.

2.3.1.2 Comparison to CVC syllables

For all comparisons that were tested, vowels in CV and VC words were significantly longer than vowels in CVC words. This is shown for all CV, VC, and CVC words below.

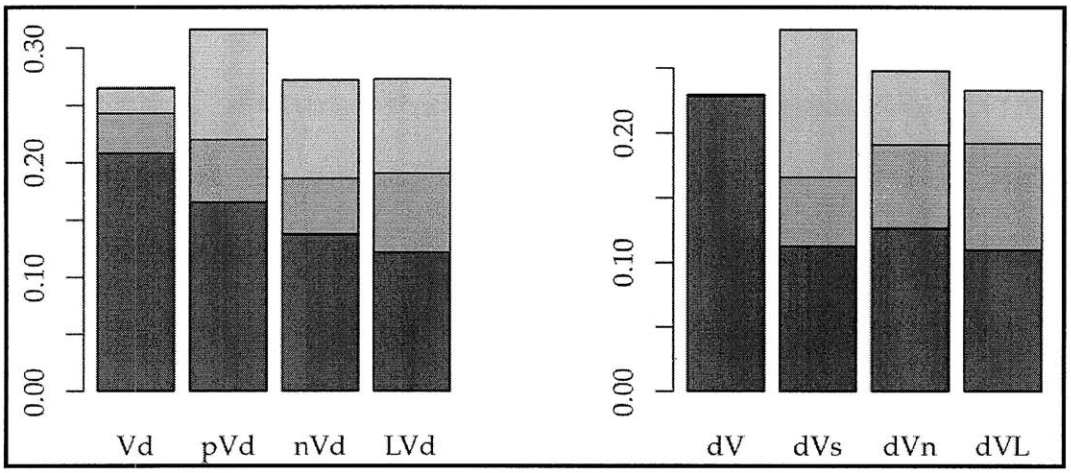

Figure 2.9. *Portion-by portion duration, in seconds, for CV, VC, and CVC syllables, separated by syllabic position of target consonants. The bottom bar represents steady-state vowel duration; the middle bar the CV or VC transition; and the top bar steady-state consonant (closure in the case of glottal stop).*

The effects ranged in size from 0.56 standard deviations (25-35 ms) for /od/ vs. /pod/ to 2.79 sds (122-175 ms) for /da/ vs. /dar/. MCMC sampling revealed that all of the tested effects were significant below the $p = 0.0001$ level.

Some complications arise in comparing the size of the simplex CS effect in various contexts. The source of the problems is the fact that the acoustic landmarks used as duration criteria for any given CVC item differ from other CVC items and differ from CV/VC items. For instance, we can ask whether simplex CS differs across vowels by comparing the [ɪd]-[lɪd], [ɑd]-[lɑd], and [od]-[lod] pairs; the answer is that the /o/ items show significantly less shortening than the other two vowels. But this could be due to several factors. In the VC items, boundary marks track the onset of voicing; in [lɪd], they track F2 maxima; in [lɑd], they track F1 maxima; and in [lod],

they track F2 minima. Furthermore, the steady-state formant values of /l/ are closer to those of /o/ than they are to the other two vowels. So any differences in steady-state duration among the pairs mentioned above may be due to differences in the relative timing of one acoustic landmark in the VC stimuli to several different landmarks in the CVC stimuli. If we could use the same landmark in all stimuli, or at least the same two landmarks in each pair (as in the [lVd]-[glVd] comparisons), we could be more certain that differences in measurement reflect actual differences in temporal coordination.

The solution adopted here is to compare CV/VC words to CVC words with /r/ as the target consonant. Boundary marks between /r/ and vowels track F3 movement in all relevant items, so landmarks are somewhat comparable across word-pairs; vowels still presumably differ in how close their F3 values are to /r/. By this test, the simplex CS effect was not significantly different across vowels, nor across onset and coda position. There was one significant 3-way interaction term: the effect was much larger in coda than in onset position for the vowel /ɑ/, relative to the vowel /o/ (46-65 ms greater difference between onset and coda for /ɑ/; $p < 0.0001$).

The magnitude of the simplex CS effect differed between subjects (although the direction of the effect did not), and adding that variation to the model resulted in a significantly better fit: $\chi^2 = 26.3$ on 2 Df; $p < 0.0001$. The size of the effect differed for various subjects by up to 34 ms from the mean effect, but all subjects showed the same qualitative pattern of simplex CS.

Subjects also differed significantly with regard to the relative size of the simplex CS effect in onset and coda position ($\chi^2 = 25.9$ on 4 Df; $p < 0.0001$). There was no significant main effect,

but the addition of by-subject random slopes reveals that three subjects had much greater shortening from VC to CVC items, two subjects had greater shortening from CV to CVC items, and one subject had essentially no difference between onset and coda (about 4% of a standard deviation). If there are differences in simplex CS depending on context, they vary in their direction and presence from subject to subject, unlike the main effect.

2.3.2 Incremental CS

Patterns of incremental CS differ by consonant quality, and they differ between onset and coda for some consonants. This is shown in the boxplot below, which compares CVC words to comparable CCVC or CVCC words.

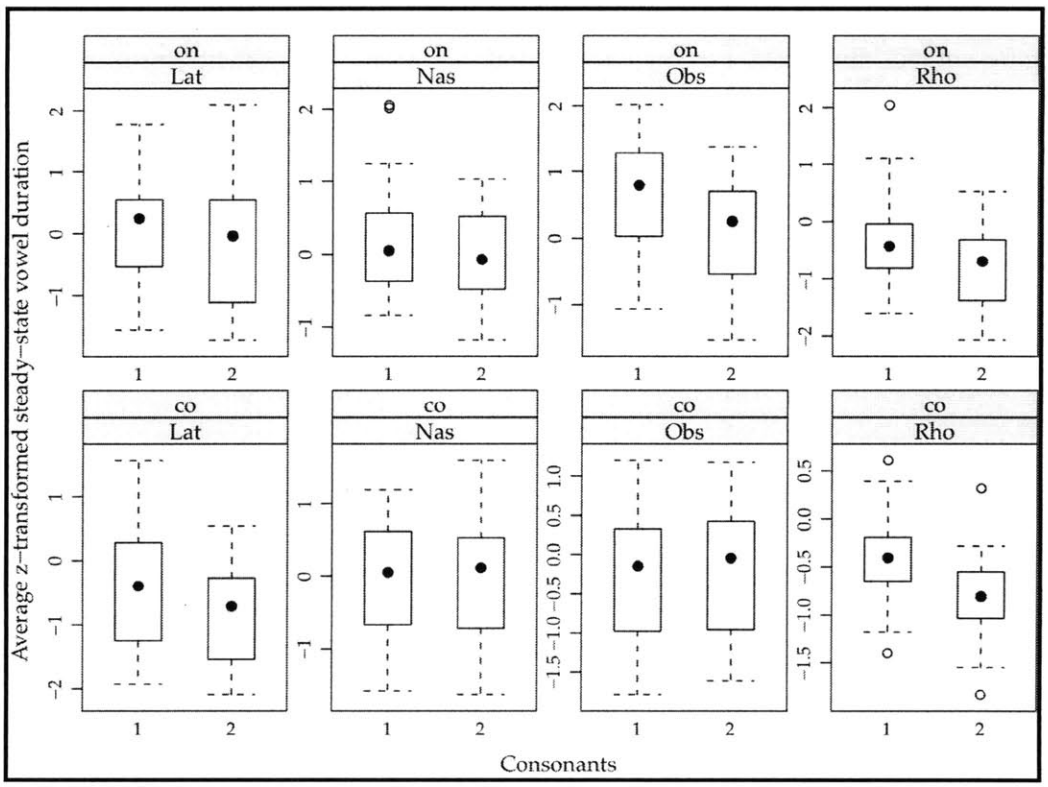

Figure 2.10. *Average steady-state vowel duration across subjects and vowels, in standard deviations from the mean. Each plot represents one manner of consonant in onset or coda position; the left bar in each plot represents duration in the singleton item, the right bar duration in the cluster item. 'Lat' = lateral, 'Nas' = nasal, 'Obs' = obstruent, 'Rho' = rhotic, 'on' = onset, 'co' = coda. For instance, the left and right bars inside the box labeled 'co' and 'Rho' show mean durations for /Vr/ and /Vrb/ items, respectively. Inside each plot, the boxes indicate the inter-quartile range (IQR), the range between the first and third quartile. The solid dot indicates the median. The whiskers indicate the range, up to 1.5 times the IQR away from the median. Open dots outside the whiskers lie more than 1.5 times the IQR away from the median and are potential outliers.*

Note that none of the interactions between incremental CS and vowel quality came out significant. This means that, broadly speaking, patterns of CS do not differ between vowels.

Including by-subject random slopes for incremental CS did not significantly improve the model: for all variables representing incremental CS effects, χ^2 statistics ranged from 2 to 9 on 7 Df; $p > 0.3$. This means that subjects did not differ with regard to incremental CS.

2.3.2.1 Liquids

Laterals and rhotics show significant incremental CS in onset position (liquids: 11-15 ms; rhotics: 16-22 ms; $p < 0.01$ for both). Both show even more incremental CS in coda position (9-13 ms more for laterals, 2-3 ms more for rhotics), but not significantly so. However, when the distinction between incremental CS with laterals and rhotics is collapsed, creating the single class 'liquids' (the difference between the two is not significant), the onset-coda asymmetry is significant: there is, on average, 8-11 ms more incremental CS in coda position; $p < 0.05$.

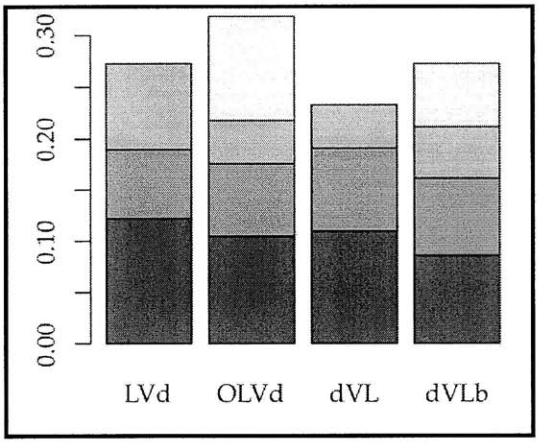

Figure 2.11. *Full duration results, in seconds, for laterals and rhotics. The bottom bar represents steady-state vowel duration; the second bar the CV or VC transition; the third bar steady-state C1; and the top bar steady-state C2.*

2.3.2.2 Nasals

Nasals show incremental CS in onset position. It is not significantly different from the amount of CS observed for laterals in onset position (1-2 ms difference between nasals and laterals). There is a small incremental CS effect for nasals in coda position (3-5 ms), which is not significant.

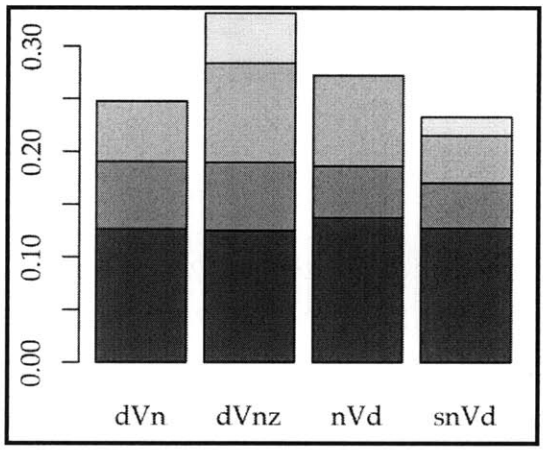

Figure 2.12. *Full duration results, in seconds, for nasals. The bottom bar represents steady-state vowel duration; the second bar the CV or VC transition; the third bar steady-state C1; and the top bar steady-state C2.*

2.3.2.3 Obstruents

/p/ in /spVd/ is followed by a shorter steady-state vowel than /p^h/in /p^hVd/. The effect is significantly larger than the onset effect for /l/ (14-20 ms larger than /l/; $p = 0.0046$). The effect is actually reversed in coda position, leading to a significant interaction between number of consonants, obstruent manner, and syllable position (29-42 ms difference between /s/ in coda

position and /p/ in onset, $p < 0.0001$). The coda anti-CS effect, 4-5 ms in magnitude, is not significant.

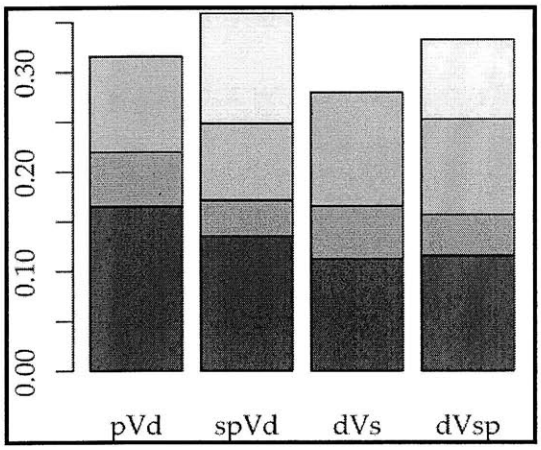

Figure 2.13. *Full duration results, in seconds, for obstruents. The bottom bar represents steady-state vowel duration; the second bar the CV or VC transition; the third bar steady-state C1; and the top bar steady-state C2.*

2.3.3 Other effects and discussion

Several other effects besides those related to the experimental hypotheses were present in the data. Words ending in /d/ had vowels that were significantly shorter when the /d/ was flapped than when it wasn't flapped (8-11 ms, $p < 0.0001$). This is presumably an effect of increased speech rate or smaller prosodic junctures, both of which could lead to shorter vowels and make flapping more likely.

Those vowels in vowel-initial words that were preceded by creaky transitions were significantly longer than those that were not (29-41 ms, $p < 0.0001$). This reflects the issues with metric m3 discussed in section 2.3.1. The results may suggest that only part of the preceding formant

transitions 'count' as vowel when compared to a word with initial glottal closure. Alternatively, they may indicate that this realization is more likely at slower speech rates than at faster ones.

There was a significant acclimation effect over the course of the experiment: vowels got shorter by about 0.03-0.04 ms in every successive item in the experiment, on average ($p < 0.0001$). This would average out to a shortening of about 3-4 ms. between successive utterances of a single item. Subjects differed in the presence/absence and magnitude of this effect: three subjects had reasonably large acclimation effects; two subjects showed effects of less than 0.02 ms/item; and one subject had a small effect in the opposite direction, which could be characterized as fatigue. Including this variation in the model significantly improved the fit: $\chi^2 = 30.2$ on 6 Df; $p < 0.0001$.

Although nearly all of the experimental items were presented as 'nonce-words', with orthography and meanings that don't correspond to existing English words, some of them are homophonous with existing English words. Neither the existence of a homophonous word nor frequency differences between existing homophones had a significant effect on steady-state vowel duration. Examination of the model shows that, while estimates of these effects were fairly large (up to 11 ms of lengthening for words with low-frequency homophones vs. no homophones and high-frequency homophones), the standard error was even larger.

As noted in section 2.2, the carrier sentence *Michael baked the ___ all in one batch* appeared to elicit a larger prosodic juncture adjacent to the target word than the other carrier sentences. Consistent with this observation, the random intercept assigned to this carrier sentence had a

higher positive value (indicating longer vowels) than all other sentences. However, the effects associated with carrier sentence were very small overall, and the estimate for this particular intercept is at most 1-2 ms of lengthening. It is possible that the normalizing fixed effects discussed above accounted for a good deal of the prosodic variation associated with differences in carrier sentence, leaving less variation for the random intercepts to account for.

2.3.4 Transition effects

A separate model investigated how the duration of the transition between vowel and adjacent consonant changes depending on syllable structure and consonant manner. /p^{h}Vd/ and /spVd/ words were excluded from this model, because their transitions (aspiration and formant transitions, respectively) are not comparable to one another.

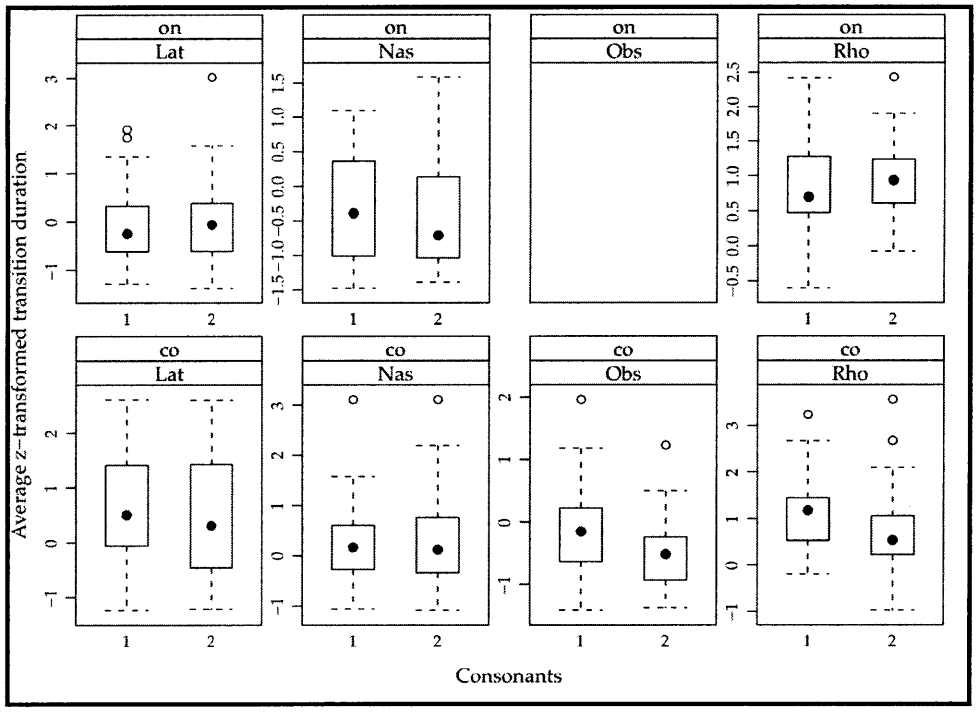

Figure 2.14. *Average transition duration across subjects, in standard deviations from the mean. Each plot represents one manner of consonant in onset or coda position; the left bar in each plot represents duration in the singleton item, the right bar duration in the cluster item.*

Transitions in CVCC words are shorter than their counterparts in CVC words by less than 2 ms on average. This effect is not significant. The (lack of a) shortening effect does not interact significantly with syllable position or vowel quality. There is one significant interaction involving consonant quality and shortening: the transitions between /s/ and the adjacent vowel show significantly more shortening from /dVs/ to /dVsp/ words than the other consonant manners show (7-11 ms. more shortening, $p = 0.0027$).

Subjects do not differ significantly for any transition effects.

Transitions have a tendency to be longer in coda position than in onset position. This effect differs by vowel, however, and is not observed for /ɑ/.

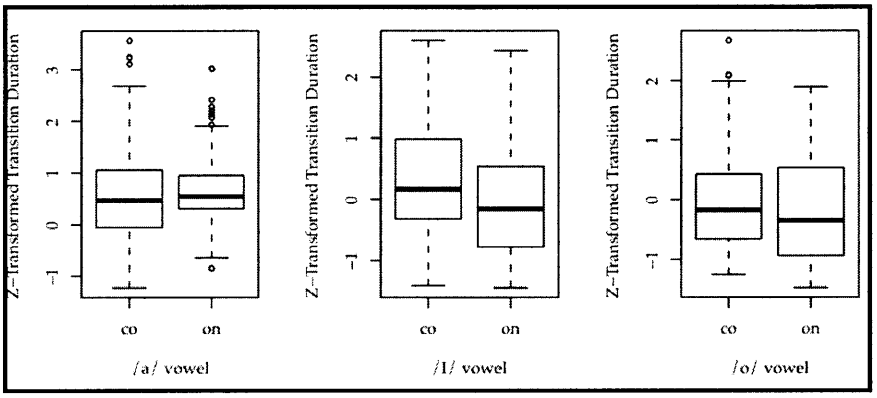

Figure 2.15. *Average transition duration across subjects and items, in standard deviations from the mean. Each plot represents one vowel; the left bar in each plot represents duration in coda, the right bar duration in onset.*

The onset/coda asymmetry is significant for /o/: 8-13 ms difference, $p = 0.002$. It is even larger for /ɪ/: 16-26 ms larger difference, $p < 0.0001$. The difference is reversed for /ɑ/: transitions are somewhat longer in onset position, but not significantly so.

Of the other effects examined, only acclimation was significant: transitions got shorter by 0.01-0.02 ms in every successive item in the experiment, on average ($p = 0.0151$). This would average out to a shortening of about 1-2 ms. between successive utterances of a single item.

2.3.6 Consonant effects

A third model investigated how the duration of the adjacent consonant changes depending on syllable structure and consonant manner. The general pattern is that there is clear shortening for all manners in onset position, while all of the manners except obstruents show a reversal to an *anti*-CS pattern in coda position. However, there are reasons to believe that the anti-CS effect may be an artifact of the segmentation strategy used; this issue is taken up in section 2.4.

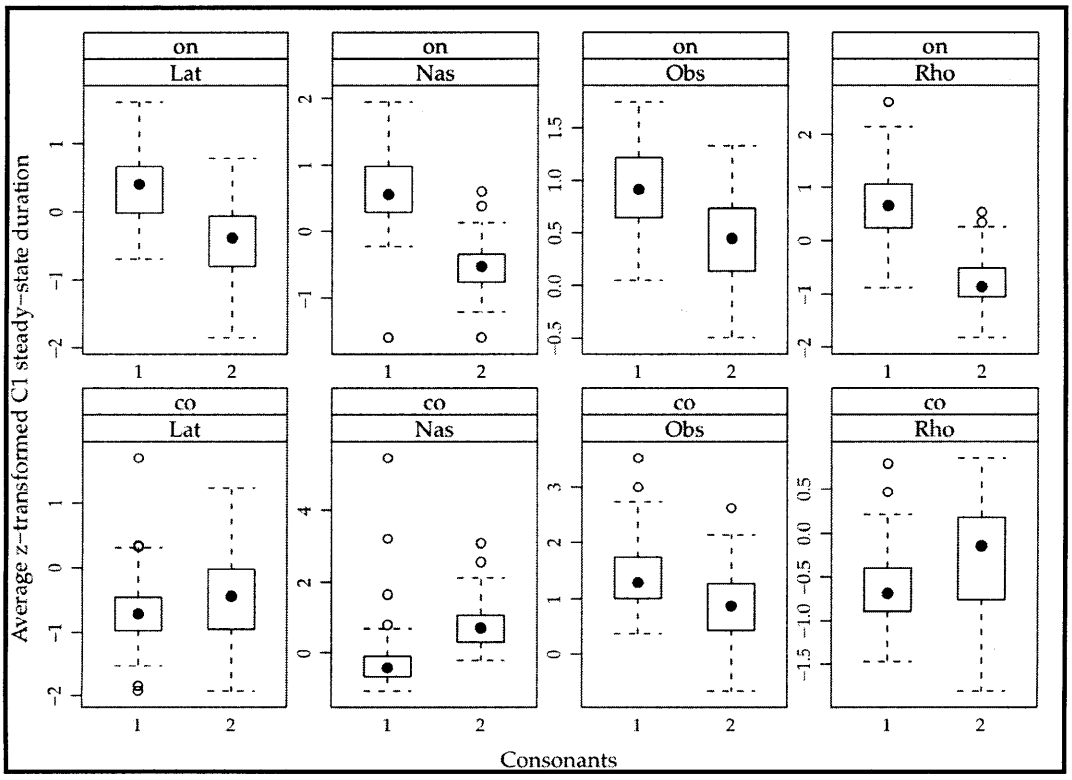

Figure 2.16. *Average C1 duration across subjects, in standard deviations from the mean. Each plot represents one manner of consonant in onset or coda position; the left bar in each plot represents duration in the singleton item, the right bar duration in the cluster item.*

2.3.6.1 Obstruents

The closure portion of /p/ is significantly shorter in /spVd/ words than in /p^hVd/ words (12-18 ms, $p < 0.0001$). The shortening effect is even larger in coda position (3-4 ms more shortening), but not significantly so. This pertains to the /s/ in /dVs/ and /dVsp/.

Subjects differ with regard to the onset/coda asymmetry, and assigning by-subject random slopes for this variable significantly improves the model fit: $\chi^2 = 9.5$ on 3 Df; $p = 0.024$. Recall that the fixed effect is non-significant; further examination shows that three subjects shorten more in coda than onset position, while the other three subjects shorten more in onset than coda position.

2.3.6.2 Liquids and nasals

/l/, /n/, and /r/ all show more shortening from CVC to CCVC than /p/, which already shows significant shortening. The difference between /p/ and /l/ is only marginally significant (6-9 ms more shortening of /l/, $p = 0.0825$). The comparisons of /n/ and /r/ to /p/ are significant: 16-37 ms more shortening of /r/ and /n/, $p < 0.0001$ for both comparisons).

/l/, /n/, and /r/ all show a reversal of the shortening effect in coda position, leading to significant 3-way interactions: $p < 0.0001$ for all comparisons. In coda, /l/ shows the least lengthening, 1-3 ms, and the effect is not significant. Rhotics show significantly more lengthening than /l/ (10-15 ms more lengthening for /r/, $p = 0.0219$). /n/ also shows significantly more lengthening than /l/ (26-40 ms more lengthening for /n/, $p < 0.0001$).

Subjects differ with regard to the magnitude of the reversal from onset to coda position for /l/. Assigning a by-subject random slope for this variable significantly improves the model fit: χ^2 = 20.1 on 5 Df; p = 0.0012. Further examination reveals that all subjects except one had a reversal from CS in onset to lengthening in coda. The lone exception, subject ME, shows substantially less shortening in coda but no actual reversal. This is shown below, with /l/ duration separated by subject and syllable position.

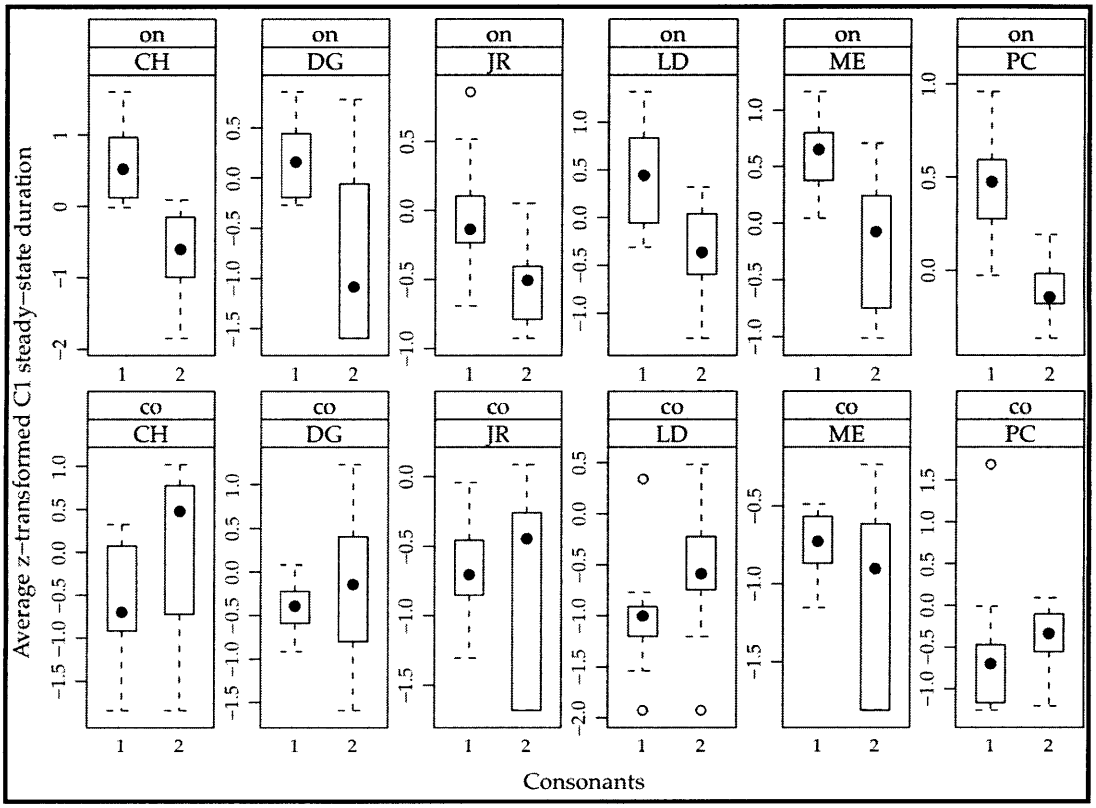

Figure 2.17. *Average /l/ duration across vowels, in standard deviations from the mean. Each plot represents onset or coda position for one subject; the left bar in each plot represents duration in the singleton item, the right bar duration in the cluster item.*

2.3.6.3 Onset/coda asymmetries

Across liquids and nasals, steady-state consonant duration tended to be shorter in coda than onset position. Obstruents were not included in this comparison because different obstruents were tested in the two positions.

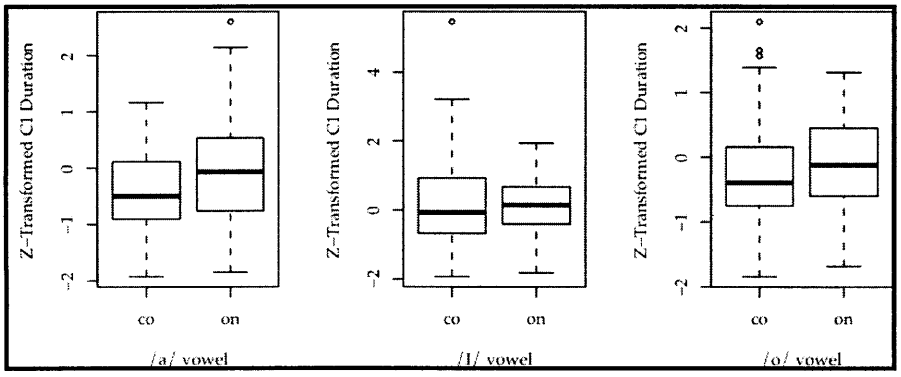

Figure 2.18. *Average C1 duration across subjects, in standard deviations from the mean. Each plot represents one vowel; the left bar in each plot represents duration in coda, the right bar duration in onset.*

The effect is significant for nasals and laterals: 28-42 ms difference; $p < 0.0001$. It is even larger for rhotics: 10-15 ms larger difference; $p = 0.006$. The effect is significantly smaller for /ɪ/ than the other vowels: 13-20 ms smaller difference; $p = 0.004$. The lone exception to this vowel asymmetry is the lateral series, where the effect is somewhat *larger* for /ɪ/, leading to a significant 3-way interaction: 5-8 ms opposite effect; $p = 0.001$ for interaction.

2.3.6.4 Other effects and discussion

There is significant interaction of syllable structure with vowel quality, specifically for /ɑ/. Incremental effects across the board tend more towards shortening with /ɑ/ than with other vowels. This means that CS is greater with /ɑ/, and anti-CS lengthening is smaller (6-9 ms, $p = 0.0007$). It's possible this has to do with the measurement criteria used for /ɑ/ as opposed to other vowels. The offset of the transition, and the onset of the steady-state consonant, was generally judged by an F1 minimum next to /ɑ/; F2 played a larger role next to /o/ and /ɪ/. If there is an asymmetry in incremental CS between F1 trajectories and F2 trajectories, it may help explain this effect. The only other explanation that comes to mind is that this effect may be due to the relatively long inherent duration of /ɑ/; if this were the case, however, we would expect compression asymmetries in more than just this one domain.

Of the other effects examined, only acclimation was significant: consonants got shorter by 0.02-0.04 ms in every successive item in the experiment, on average ($p < 0.0001$). This would average out to a shortening of about 2-4 ms. between successive utterances of a single item.

2.3.7 Summary of results

The preceding sections have enumerated a long list of results. In this section, we collect and summarize those results. We further distinguish between results that seem to reflect general

compression effects and effects that can be explained by artifacts of the segmentation strategy or the particular materials elicited in the experiment.

All consonants drive simplex CS in both onset and coda position. The statistical model finds some asymmetries by manner, vowel, and position, but these are confounded by differences in the acoustic criteria used for segmentation in various contexts. Examining the segment with the most consistent criteria across contexts, /r/, there does not appear to be significantly greater coda CS than onset CS.

The incremental CS results for steady-state vowels, on a first pass, are summarized as follows:

Incremental CS

	Onset	Coda
Obstruent	Y	N
Nasal	Y	N
Liquid	Y	Y

Table 2.3. *Presence of incremental CS effect for steady-state vowel as a function of C1 manner and syllable position; measurement criterion excludes formant transitions of /sp/ clusters.*

Recall, however, that the onset obstruent items are not a perfect comparison; it is not obvious what the best comparison is. While the period of *steady-state* vowel is shorter in /spVd/ than in /p^{h}Vd/, /spVd/ also contains a period of (modally-voiced) formant transitions into the vowel that /p^{h}Vd/ does not. When that period is taken into account, there is a marginally significant *anti*-CS effect (6-9 ms, $p = 0.079$). Because aspirated stops don't occur as the second consonant in English clusters, and voiceless unaspirated stops don't occur as singleton onsets, this is the best

comparison we can manage, but it is not a straightforward singleton-cluster pair. Comparing the duration of modally-voiced vowels, the table looks as follows:

Incremental CS

	Onset	Coda
Obstruent	N	N
Nasal	Y	N
Liquid	Y	Y

Table 2.4. *Presence of incremental CS effect for steady-state vowel as a function of C1 manner and syllable position; measurement criterion includes formant transitions of /sp/ clusters.*

Not reflected in the table is one finding about the magnitude of incremental CS: liquids induce a slightly larger incremental shortening effect in coda than in onset position, particularly /l/.

For vowel steady-states, incremental CS effects did not vary by subject nor by vowel.

For transitions and consonant steady states, results were somewhat more variable. In general, transitions do not shorten between singleton and cluster words. There is one exception to this, /dVs/-/dVsp/, which is discussed in the next section. Consonant steady states do display CS in onset position. There is an anti-CS lengthening effect in coda position, but we argue in the next section that this is confounded by other differences between the CVC and CVCC stimuli.

Transitions tend to be generally longer in coda position than in onset position, while consonant steady-states show the opposite pattern. This might be taken as evidence for a trading relationship between consonant duration and transition duration, which is predicted by both of the approaches outlined in the introduction. Both patterns show idiosyncratic reversals across

vowels however, and both should be interpreted with caution. The design of the experiment does not allow for completely and strictly controlled comparisons between boundaries in onset and coda position; indeed, there may not exist such a design.

2.4 Discussion and conclusions

2.4.1 Distinguishing between the articulatory and perceptual models

The study finds that simplex CS for steady-state vowels is present in both onset and coda positions; incremental CS for vowels is induced by liquids in both positions, nasals in onset, and is not clearly present for obstruent sequences in either position. Incremental consonant shortening affects all consonants in onset position (regardless of vowel shortening); coda consonants are discussed below. Transitions between consonants and vowels do not shorten, with one exception discussed below.

In section 1, we developed schematic predictions about CS based on two broad theoretical approaches, one articulatory in nature and one auditory. We now turn to the question of how well each of the approaches can account for these results.

The articulatory approach makes more specific and more uniform predictions, and therefore is the stronger hypothesis, to be preferred *a priori*. Those predictions appear to be too uniform to account for the data, however. The version of the articulatory theory that includes asymmetries in the presence of a C-center effect predicts that compression effects should be driven by onset consonants but not coda consonants. This is falsified by the results of the study.

One might attempt to modify the theory to accomodate the results. One possibility is that there really is a C-center effect present in coda position, but something about the variability of coordination between vowels and coda consonants has made it difficult to detect in articulatory studies. This essentially turns the C-center version of the theory into the simpler Fowler theory. This hypothesis, however, can't explain why the incremental coda effects are limited to items with liquids.

One further modification might try to explain *that* asymmetry as well. Gestural investigations generally find that, in coda position, the vowel-like dorsal gesture associated with /l/ precedes the tongue-tip gesture, impinging on the preceding vowel (Sproat & Fujimura 1993, Browman & Goldstein 1995, Proctor 2009 for a review). If this dorsal gesture is half-way in between a vowel and a consonant, it might display some kind of mixed behavior, impinging on the preceding vowel in cluster-like fashion while also being repelled incrementally from following consonants. Even if this could be formally worked out, however, it would not explain the data. For one thing, it is not clear whether English rhotics display a similar articulatory asymmetry. Furthermore, nasals do display a similar coda asymmetry, whereby their velar abduction gesture is phased earlier with respect to their oral constriction gesture. If the asymmetry in CS is to be explained by the gestural properties of wide as opposed to narrow constrictions, it will need to somehow connect the conditioning of CS to the difference between /n/ and /l/ in this regard; they do not condition identical patterns of incremental CS.

The patterns of CS discovered in this study do not seem to be amenable to explanation in articulatory terms. What, then, of the auditory theory? That theory predicted that compression

should be observed in codas, but not necessarily in onsets. It also predicted that there could be asymmetries between various segments in the conditioning of CS, which would be based on the auditory properties of those segments.

Clearly, onset consonants do drive CS. This result replicates several previous studies described in section 1. This poses problems for the analysis of vowel-length contrasts and their distribution, which are briefly discussed in the next section.

Apart from this, the auditory theory seems capable of explaining most of the results, when coupled with specific hypotheses about recoverability. We briefly outline some of those speculative hypotheses here; they will be investigated in detail and tested empirically in the following chapters.

The hypothesis formalized in the next chaper and explored in the next several chapters is that larger vowel-compression effects are observed in syllables that include higher-sonority segments adjacent to the vowel, because higher-sonority segments allow more information about that vowel to be recovered.[6] To explain all of the asymmetries observed here will also require a minimum inherent 'floor' duration for vowels, as mentioned in section 2.1.

To illustrate the logic of the proposal, we first consider the variation in VC items discussed in section 2.3.1. We saw in section 2.3.1 that the particular phonetic realization of VC syllables correlates with vowel duration. Some of these tokens are produced with a glottal stop preceding

[6] In fact, the relevant notion here is not exactly sonority, but something like 'transparency with respect to the features of an adjacent vowel'. By hypothesis, the two notions correlate in English.

the initial vowel, some with a modally-voiced glide transition from the vowel in preceding *the*; most tokens feature a realization somewhere between these two extremes. The duration of steady-state vowel is shorter in tokens that are preceded by formant transitions.

When we introduce a duration metric that counts these transitions as part of the vowel, they instead come out *longer* than comparable tokens without formant transitions. This can be analyzed and explained in a model where vowel duration is not simply a property of acoustic steady states, but may be dispersed over different parts of the acoustic signal. In this approach, a vowel's effective duration is associated with its recoverability: parts of the signal that contain steady-state vowel obviously contribute a lot to perceptibility; adjacent parts of the signal that are affected by the vowel may also contribute to the vowel's perceptibility, and may therefore be perceived as part of the vowel's effective duration.

Glottal closure conveys little or no information about the following vowel; formant transitions convey a lot of information. In the glottal stop realizations, then, that following vowel will need to be relatively long in order to convey as much information as the combined steady state and transitions convey in the realizations with transition. The crucial idea in this approach is that vowels have a *target* for something like recoverability over time, rather than simple duration. Steady state duration, of course, will help fulfill that target; other portions of the signal will also help fulfill the target, in proportion to how informative they are about vowel quality.

This approach can be extended to account for most of the data in the experiment, when coupled with assumptions about the relative perceptual informativity (for vowel quality) of various

portions of the signal. Those assumptions will be more fully elaborated in chapter 3, and tested in chapter 4.

One broad asymmetry encountered here is that incremental CS is observed in comparisons of items containing liquids adjacent to the vowel, but not items that contain only obstruents. If the recoverability hypothesis is correct, this asymmetry must hold because either the liquid steady state or the transition between vowel and liquid (or both) conveys more information about the adjacent vowel than the comparable intervals do in obstruent items. This seems plausible at a first pass: liquids and their transitions have clear formant structure that could change based on an adjacent vowel; obstruents are realized acoustically as some combination of noise and silence. Silence, obviously, conveys nothing about an adjacent vowel; noise should change somewhat depending on the vowel context, but our hypothesis predicts that this variability is less informative about vowel quality than variability in liquids is.

Even given this asymmetry, we might predict that obstruents induce less incremental CS for vowels, but we would still predict some. One possibility is that there really is a small effect, but the current study is not precise enough to uncover it; perhaps the effect is tiny in comparison to between-subject effects or random noise introduced by a failure to perfectly control for prosodic factors. In this case, there would be nothing left to explain. The more prudent response, however, would be to assume that the lack of incremental CS is real, and ask how it might be explained.

The hypothesis explored in the next chapter is that in some cases, the recoverability or effective duration of the adjacent vowel hits a 'floor' after adding just one (low-sonority) segment; in

these cases, adding further segments (as in CVC vs. CVCC) will not result in further shortening. This is why some consonants in some positions do not drive incremental CS.

A further asymmetry concerns items with a nasal as C1; they display incremental CS in onset position, but not coda. The recoverability hypothesis can explain this asymmetry if something about coda nasals makes them less informative than onset nasals with regard to vowel contrasts. Again, this entails that either the consonant steady-state, the transition to vowel, or both carry less information about an adjacent vowel in coda than in onset position. One plausible property is the amount or extent of nasalization overlapping the adjacent vowel.

The velar abduction (by lowering) gesture for nasals is 'stronger' in coda position than in onset, in several senses. According to Krakow (1999), "[t]he larger velum lowering movement, lower minimum and longer low plateau indicate that a vowel preceding a [syllable- or word-]final nasal is more likely to be affected by coarticulatory nasality than a vowel preceding an [syllable- or word-]initial nasal." All else being equal, nasality during the preceding vowel or transition will make vowel contrasts less distinct (Wright 1986, Beddor 1993). If 'duration' targets are actually recoverability targets, we can explain the asymmetry in incremental CS for nasals.

One might object that onset /sn/ and coda /nz/ are not comparable in the first place, because the outer consonants differ in voicing and the /nz/ sequence is likely to be interpreted as a morphologically complex plural noun. We digress for a moment to argue that these hypotheses are not likely to explain the data.

As to voicing, the hypothesis might be that vowel-lengthening associated with a final voiced obstruent could negate the default pattern of incremental CS. The coda liquid clusters, however, also contained a voiced stop in final position, but incremental CS was observed for these items.

The morphological confound would rely on the idea that, in spite of the orthography (the words were not written with /s/), subjects analyzed items ending in /nz/ as morphologically complex. Given that assumption, we might then hypothesize that compression effects only hold internal to a morpheme, and that this is why no incremental CS was observed in these items. We can't conclusively rule out this explanation, although /nz/ is available morpheme-internally in English: examples include *lens*; *cleanse*; the colloquial use of the proper name *Jones (for)* as a singular noun or verb meaning 'desire'; the colloquial use of the proper name *Benz* as a singular noun meaning 'Mercedes Benz car'; and the neuroanatomical term *pons*.

Returning to the asymmetries found in the experiment, we now consider liquids. They show a similar asymmetry to nasals, but in the opposite direction: driving greater CS in coda position than in onset. The recoverability hypothesis can explain this asymmetry if something about coda /l/ and /r/ make them more informative about vowel quality than onset /l/ and /r/. As mentioned above, Sproat & Fujimura (1993) find that the relative timing of the tongue-tip constriction gesture and the tongue-body constriction gesture involved in /l/ changes from onset to coda position. In onset position, the two gestures are more or less simultaneous, reaching peak displacement at roughly the same time; in coda position, however, the tongue body gesture leads, with the tongue-tip gesture starting around the time of peak tongue-body displacement. What this means for the recoverability hypothesis is that the l-V transition in onset position consists of both

a tongue-tip and tongue-body gesture overlaid on or blended with the following vowel gesture, while the V-l transition in coda position consists of only the tongue-body gesture overlaid on or blended with the vowel gesture. All else being equal, more obscuring gestures should result in inferior recoverability; this could explain the syllable-position asymmetry.

In addition, the *stiffness*, velocity, and degree of constriction are lower for the /l/ tongue-tip gesture in coda than in onset position. This suggests that some of the asymmetry may also be attributed to the characteristics of the /l/ steady-state, although that steady-state also tends to be shorter in coda position.

Less is known about the timing of various articulatory gestures involved in English /r/. This segment can include at least three gestures: a tongue-tip or -blade constriction, pharyngeal constriction, and rounding or protrusion of the lips (Alwan *et al.* 1997). The tongue gestures, at least, show a fair amount of variability between subjects and contexts, including trading relationships (Alwan *et al.* 1997, Guenther *et al.* 1999). If English /r/ patterns with /l/ and /n/ in initiating its wider constriction gesture earlier in coda than onset position, then the CS asymmetry is explicable in exactly the same terms as /l/. Although I am not aware of any research on this point, it appears to be the case that at least for Spanish /r/ in coda position, tongue-body activity precedes tongue-tip activity (Proctor 2009). Of course, this segment differs from English /r/ in many respects.

One final point that may be explicable in terms of perceptual asymmetries is the comparison between onset /p^h/ and /sp/ sequences. This study found that the steady-state vowel following

/sp/ is shorter than the modally-voiced vowel following /p^h/. This could be explained if the formant transitions adjacent to /sp/ contain more information about a vowel than aspiration does, or if the presence of the /s/ offers an advantage, or both. Another possible explanation, however, might be that /sp/ is simply longer than /p^h/, hence induces more shortening.

2.4.2 Some effects that do not have a perceptual explanation

There were no shortening effects observed for transitions in any of the consonant series except one. This is consistent with a model in which acoustic transitions are basically determined by interpolation between steady-state targets, and are not actively manipulated (for durational properties) by the speaker. The lone exception is the coda /s/-/sp/ comparison, where V-s transitions shorten significantly in /Vsp/ words. Closer examination of the relevant materials suggests that this is due to a difference in timing between events internal to the segment /s/. Specifically, transitions in /Vs/ words are often marked by formant movements beginning well before breathiness and/or an abrupt decrease in energy above the first formant; in /Vsp/ words, the two changes tend to begin closer to the same time. Illustrative tokens are shown below.

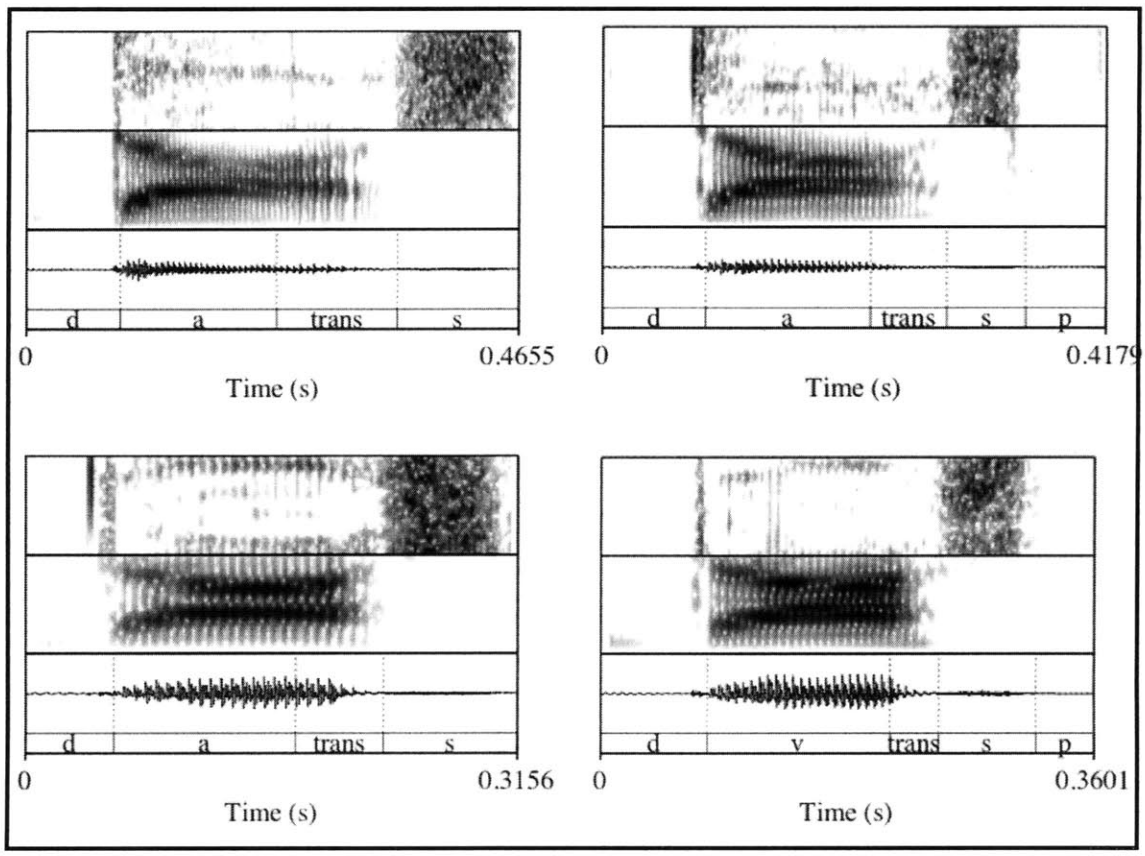

Figure 2.19. *Illustrative examples of /dɑs/-/dɑsp/ for speakers CH (top) and DG (bottom). Transitions shorten in /dɑsp/ tokens relative to /dɑs/. Spectrograms show F1 and F2 (bottom) and noise in the 5.5-10 kHz range (top).*

The finding is that the portion of signal from onset of formant movements at the end of the vowel (a rise in F2 in this case) to onset of the /s/ steady state (a peak in energy above 5 kHz) is shorter in /dVsp/ tokens than /dVs/. This is related to another observation: the onset of attenuation of energy above F1 relative to the onset of formant transitions occurs earlier in /dVsp/ tokens than /dVs/; this is why the rise in F2 is much easier to see in the /dɑs/ tokens above. Transitions in the

/dVs/ case consist of first formant movement than attenuation of energy; transitions in the /dVsp/ case consist of both changes at once.

This raises the question of *why* the timing changes. One speculative explanation concerns the shortening of steady-state /s/ in cluster tokens relative to singleton ones. If we assume that the attenuation of energy is related primarily to a glottal abduction gesture, while changes in F2 are related primarily to a tongue-tip gesture, it suggests that the tongue-tip gesture precedes the glottal gesture in singleton tokens, but the two are closer to simultaneous in cluster tokens. This would follow naturally from the shortening of the /s/ if the glottal gesture is timed to coincide with some point internal to the /s/.

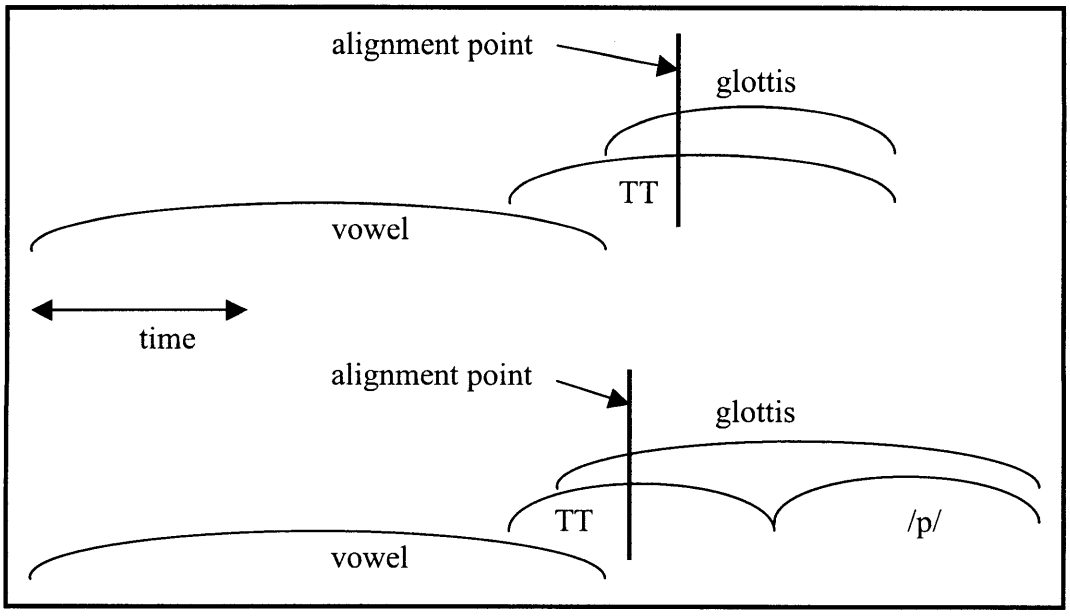

Figure 2.20. *Illustration of how articulatory shortening of /s/ might result in a shorter acoustic transition for /dVsp/ tokens (bottom) than /dVs/ tokens (top). TT = tongue-tip gesture.*

Figure 2.20 shows how /s/ shortening might lead to transition shortening. In this scenario, the glottal gesture bears a constant temporal relationship to the attainment of constriction target for the tongue-tip gesture. As the tongue-tip gesture shortens, target attainment occurs earlier relative to the vowel (this assumption only holds for compression, not truncation). Assuming that the acoustic steady state of /s/ occurs at a fixed point relative to the glottal gesture, the measured acoustic transition will be shorter in the cluster case. This scenario is meant only as a suggestion of how the observed timing pattern may arise. We can't know whether this it is correct in the absence of articulatory data that bear on the issue.

The study also found unexpected patterns of anti-CS lengthening for /r/, /n/, and a small (non-significant) effect for /l/ in coda position; for instance, the steady state of /r/ is longer in /dorb/ than in /dor/. This is most likely a result of the fact that the criteria for marking the offset of these segments in CVC and CVCC syllables are not strictly comparable. In CVC syllables, the offset was marked as soon as the following preposition-initial vowel began to influence the acoustic signal. In CVCC stimuli, the offset was marked where the following obstruent abruptly changed the acoustic signal. The difference in duration between CVC and CVCC syllables can be explained under the hypothesis that a following vowel manifests itself acoustically earlier in the sonorant steady-state than a following obstruent does.

2.4.3 Comparison with previous studies

The majority of the results that can be directly compared to previous studies replicate those studies. However, certain findings seem to contradict earlier studies. In this section we attempt to explain the discrepancies.

The finding that obstruents do not condition incremental coda CS is in direct conflict with the findings of Munhall *et al.* (1992): they found that there is a small incremental coda CS effect for obstruent clusters. The current results may also be taken as a contradiction of Fowler's (1983) similar findings; that study, however, reported no statistical analysis and made no distinctions by consonant manner.

There are several possible reasons why the findings might differ. One possibility is that the measurements are different in the various experiments. The current study did find that the VC transition shortened in these cases, as discussed above; if part of this interval was included in the vowel measurement in the other studies (and the descriptions in the Munhall *et al.* paper suggest this is probably the case), it would result in a measured incremental CS effect. It is unknown whether this explanation will generalize to stop-fricative clusters, however, as the current experiment only examined /Vsp/ items.

A second possibility is that the speech elicited in the other studies was different from the speech elicited here. In particular, it was probably more rhythmically constrained, due to the repetition of a single carrier sentence in the Munhall *et al.* study, and the repetition of words to a

metronome beat in the Fowler study. In chapter 3, I demonstrate how an extra-linguistic, task-specific constraint enforcing isochrony might lead to extra compression effects beyond those encountered in normal speech.

A third possibility is that the subjects in the other studies did indeed show the relevant effect, but the effect doesn't generalize to the population of English speakers. Fowler's data is based only on utterances from the author. Munhall *et al.* test three speakers, but perform separate analyses of variance for each subject. Neither of these procedures allows one to generalize results to the broader population of English speakers; the statistical issues are discussed in length by Max & Onghena (1999). In essence, this explanation says that the effect was observed 'by accident', and if we had observed more speakers it would likely have averaged out to 0. This is always a danger when we conduct studies with very few subjects.

A final possibility, noted briefly in section 2.4.1, is that there is a very small incremental effect in obstruent clusters that the current study was unable to detect. This is certainly a plausible explanation. The effects in the Munhall *et al.* study are generally rather small (the largest is 36 ms but most comparisons are on the order of 3-10 ms) and vary between subjects. Standard deviations are also extremely small: reconstructing from the standard error terms given in the paper, they seem to be on the order of 5-25 ms. This is a fraction of the variance observed in the current experiment, which would make small effects easier to detect. This difference in variance is presumably due to the rhythmic factors mentioned above. This explanation, although plausible, is less conservative than the isochrony explanation, because it assumes there is no

difference in speakers' behavior rather than trying to explain the differences that appear to exist. For this reason, we pursue the isochrony explanation in the next chapter.

The series of studies by Crystal & House (1982 *et seq.*) contradict the findings of the current study and all comparable studies described in this chapter. For instance, they find no strong evidence for compression effects in stressed syllables, and report that the sonorant/obstruent distinction has no effect on the duration of a preceding vowel. These studies, however, suffer from a host of methodological and analytical defects.

First, the authors tend not to control their data for consonantal or vocalic features, nor for the identity of the speaker. The 1988 paper gives more detail about the context of some of these data, but consonantal and vocalic features are never specified at the same time. If CS effects are smaller than inherent duration differences between segments, speech rate effects, cross-subject differences, or if they differ across various consonants in any way, they wouldn't be visible in these data. Van Santen (1992) criticizes this work on these grounds, illustrating his point with several examples of how factor-confounding led Crystal & House to posit spurious effects.

Second, the authors repeatedly state that segmental boundaries were marked using 'standard criteria' with regard to the speech waveform or the spectrogram, without further specification of what those criteria are. Because there *is* no unique set of standard criteria for placing boundaries even between vowels and obstruents (much less vowels and sonorants), it is impossible to determine how boundaries were marked in these studies. We can make the charitable assumption that the boundary criteria were at least *consistent* between strings containing the same type of

segments; but given that the authors collapse data across segment types, and that the data are not necessarily balanced for segment type, we have no idea how to interpret their results.

Finally, none of the claims about CS in these papers are presented with statistical analysis. Most of the data is not presented in sufficient detail to test any hypotheses about CS. In the 1990 paper, for instance, the authors claim that there is no evidence for compression effects in their corpus. In a comparison of CVC and CVCC syllables, however, the conditions are neither balanced nor specified for vocalic or consonantal features. Some of the conditions do seem to show CS effects, but not all of them; this situation is apparently interpreted as an absence of evidence for any clear pattern.

The point of this critique is not to be dismissive of these researchers' efforts. There are often valuable insights to be gained from exploratory and descriptive studies, and Crystal & House undoubtedly uncovered valuable patterns; compression efffects were not the main concern of their investigation. The argument is simply that a range of other studies with various methodological and analytical advantages over these ones have found that compression effects do exist, and that descriptive statistics can not be evaluated on an equal footing with rigorous, controlled analyses using inferential statistical tests.

2.4.4 The trouble with vowel-length neutralization

As noted earlier, the current study and several previous ones have established that onset consonants do drive vowel shortening in at least English, Swedish, and Dutch. If neutralization

patterns occur because of shortening, we should find languages where neutralization occurs in syllables with onsets, but not syllables without onsets. We know of no such language; the licensing of vowel-length contrasts does not appear to interact with the presence of a preceding consonant. This means that the relationship between vowel-length contrasts and compression effects is not fully understood.

A more conclusive answer to the question of what governs the distribution of vowel-length contrasts can presumably only be answered through detailed studies of the timing and phonology of many languages. Nonetheless, we offer some speculation here. It is possible that the perceptual distinctiveness of vowel-length contrasts is affected less by the absolute amount of duration available to express the contrast than it is by the perceptual *sharpness* of the vowel boundaries. This explanation might help us to explain the generalization that licensing of the contrast sometimes interacts with the presence of a following consonant, but does not appear to interact with the presence of a preceding consonant. The current study found that transitions between a vowel and coda consonant are longer than those between a vowel and onset consonant, while the duration of the steady-state consonant varies inversely with the transition. This seems to agree with the articulatory observation, mentioned above, that coda consonants feature constrictions that are weaker in several senses. The perceptual consequence of this could plausibly be that coda consonants blend more with an adjacent vowel than onset consonants do. Making the further assumption that articulatory and acoustic blending are antagonistic to detecting boundaries, we could derive the prediction that the distinctiveness of vowel length contrasts will be damaged more by a following consonant than by a preceding one.

This hypothesis would predict a hierarchy of contexts where neutralization is more or less likely, based on the relative crispness of the vowel-consonant boundary. It is not clear that such a rich typology exists; most of the examples we know of simply concern the presence or absence of a consonant.

Another possibility is that other languages either fail to display onset-driven CS at all or have an effect much smaller than that driven by coda consonants. The only languages in which onset-driven CS is attested, as far as we are aware, are English, Dutch, and Swedish. These represent a relatively narrow range of (Germanic) languages. If other languages tend to feature only coda-driven CS, it would explain why onsets do not appear to interact with vowel-length contrasts. To know whether this is a plausible explanation, of course, would require detailed acoustic studies of languages with and without neutralization patterns.

Most of the explanations of temporal patterning observed in this study have made recourse to hypothesized perceptual differences between segments or transitions. Specifically, it was hypothesized that segments may shorten more when their surrounding context contains more perceptual cues pertaining to their presence or quality. The next chapter asks what type of grammar could predict the qualitative patterns of compression observed in this experiment. That grammar, in turn, will require qualitative assumptions about the relative perceptual properties of different types of speech event. Chapter 4 reports an experiment that attempts to test those assumptions.

Appendix 2A

Segmentation criteria used in the study

The table below lists the acoustic criteria used for segmentation. The columns represent the boundaries between vowel proper and transition, transition and C1, C1 and C2 in cluster items, and C1 and the adjacent word in singleton items, respectively. Abbreviations are *high plateau* (HP), *low plateau* (LP), *onset* (on), *offset* (off), *abrupt rise in energy above the 1st formant* (ER), *abrupt drop in energy above the 1st formant* (ED).

		Boundary			
		V-trans	**trans-C**	**C-C**	**C-#**
segment / context	l/_ɑ	F1 HP-on	F1 LP-off, ER	ER	F1 LP-on, ED
	l/_ɪ	F2 HP-on	F2 LP-off or F1 LP-off, ER		
	l/_o	F1+F2 HP-on	F1+F2 LP-off, ER		
	l/ɑ_	F1 HP-off, ED	F1 LP-off	ED	F1 LP-off or F2 LP-off
	l/ɪ_	F2 HP-off, ED	F2 LP-off		F1 LP-off
	l/o_				
	r/_ɑ	F3 HP-on	F3 LP-off	ER	F3 LP-on
	r/_ɪ				
	r/_o				
	r/ɑ_			ED	F3 LP-off
	r/ɪ_				
	r/o_				
	n/_ɑ	F2 LP-on	F1 LP-off, ER	End of silence or onset of voicing	F1 LP-on, ED
	n/_ɪ	F2 HP-on			
	n/_o	F1 HP-on			
	n/ɑ_	F2 LP-off	F1 LP-on, ED	1st appearance of aperiodic noise, ED	F1 LP-off, ER
	n/ɪ_	F2 HP-off			
	n/o_	F2 LP-off			
	p^h/_ɑ	Onset of energy around F1	Onset of aperiodic noise following burst	--	ED
	p^h/_ɪ				
	p^h/_o				
	p/_ɑ	F1 HP-on	Onset of energy around F1	Offset of HP of energy above 5 kHz	--
	p/_ɪ	F2 HP-on			
	p/_o	F1 HP-on or F2 LP-on or F2 HP-on			
	s/ɑ_	F2 LP-off	HP of energy above 5 kHz		Offset of HP of energy above 5 kHz
	s/ɪ_	F2 HP-off			
	s/o_	F2 LP-off			
	d/#_	--	Onset of energy around F1	--	ED
	d/_#				
	ɾ/_#		ED, F1-3 HP-off		ER, F1-3 LP-off

3 A constraint-based account of English CS

3.1 Introduction

In the preceding chapter, we found that adding complexity to a syllable results in compression of the units within the syllable, but not all sub-syllabic units pattern identically with regard to this compression. The question to be addressed in this chapter is what kind of a grammar might produce such patterns as outputs. We develop a formal system of weighted, violable constraints that produces qualitative patterns similar to those observed in the production study. In this account, the presence of compression effects is due to the presence of conflicting constraints on duration in the grammar. Those effects are sometimes absent when segments shorten to a point that might jeopardize their recoverability. Asymmetries with regard to compression will be attributed to hypothesized differences in the perceptual properties of various segments and transitions.

The principle findings from the production study in chapter 2 are that all segments induce simplex CS (e.g. /od/ vs. /nod/), but only some induce incremental CS (e.g. /nod/ vs. /snod/). Liquids condition incremental CS in both onset and coda position, nasals do so only in onset position, and obstruents don't clearly induce incremental CS in onset or in coda position. In addition, the amount of incremental CS for items with liquids as the inner consonant appears to be greater in coda than in onset position, especially for /l/. We proposed that these asymmetries are related to the perceptual properties of segments and transitions: vowels shorten more when their auditory features are more perceptible in the adjacent sounds.

The theoretical framework developed here accounts for patterns of CS by positing duration targets for larger units as well as smaller units within them. For this chapter and much of what follows, we will assume that the relevant larger units are syllables. This is not a logical necessity; all of the items analyzed here could equally well result from targets for prosodic feet or words, intervals from one vowel to the next, or combinations of any of these constituents. Only careful investigation will reveal which of these alternatives is correct. The goal here is to delineate what types of patterns emerge in monosyllabic content words; the production experiment was not designed to test for differences among larger constituents. The smaller units here are taken to be segments. Again, one could imagine other possible ways of segmenting the speech stream, but segments will serve our purposes for this analysis.

The two types of duration constraint come into conflict as more complexity is added to a syllable; the phonetic realization of a given string involves a tradeoff between the competing constraints. A useful metaphor for understanding the logic of the formalism is the problem of trying to fit partially-malleable objects into a partially-malleable container. Physically, each object and the container has some inherent volume, the size that it possesses when not acted upon by external forces. If the objects to be fit are larger in aggregate than the volume of the container, then all entities will depart from their inherent volumes to some extent: the objects will compress and the container will expand. The exact tradeoff in how much each entity is deformed will be determined by the relative rigidity of the objects and the container.

In this metaphor, the inherent volume of physical entities corresponds to duration targets for linguistic objects; these are the dimensions that entities 'want' to have. The rigidity of those entities corresponds to the weighting of constraints, which mediates the ways in which deviations

from targets trade off against one another.

The idea that compression effects are due to pressure from higher-level duration targets is not new, but has rarely been explicitly formalized. Lindblom & Rapp (1973) mention the idea that duration trading between segments may be a way to keep syllable duration relatively consistent, although their formal approach to compression in Swedish is rather different from the one developed here. Fujimura (1987) proposes a general timing model along these lines, using the metaphor of nested spring systems rather than nested duration targets or stuffing objects into a container. He is not explicitly concerned with compression effects, but their existence is a natural prediction of his model. Formally, his system is similar or identical to a model with weighted, gradiently-violable constraints on segments and syllables. Clements & Hertz (1996) propose a model of timing where the syllabic nucleus (in an extended sense that includes transitions and liquids) is assigned a base duration and this higher-level target constrains the durations of segments internal to the nucleus, creating trading relations. The framework developed in this chapter has similarities to each of these earlier models, but will be narrowly focused on accounting for compression effects, and will be formalized rather differently.

The phenomenon of *closed-syllable vowel shortening*, a specific form of CS discussed in chapter 2, has received a lot of attention in the phonetic and phonological literature. In this phenomenon, which is widely attested cross-linguistically, vowels in closed syllables are observed to be shorter than vowels in open syllables. Analyses of CSVS often make reference to the idea that vowel compression is due to higher-level duration constraints, on a syllable, rime, or mora. Maddieson (1985), after reviewing cross-linguistic evidence for vowel-shortening preceding geminate as

opposed to singleton consonants, suggests that the phenomenon may itself be an argument for treating the syllable rime as a unit of timing. This implicitly suggests that compression effects are due to higher-level duration targets conflicting with lower-level ones. Myers (1987) invokes this trading approach in a phonological analysis of English vowel shortening phenomena: he argues that the particular English rules are phonologizations of a universal phonetic tendency towards CSVS, and that the universality of CSVS itself follows from the fact that the syllable is a unit of timing. Flemming (2001) is a more formal approach to CSVS in this vein. He develops a model with weighted, gradient constraints to characterize the competing pressures on segment and syllable durations. This is the general framework that we adopt here.

The constraint-based framework developed here builds on ideas developed elsewhere. It is similar to Optimality Theory (henceforth OT, Prince & Smolensky 1993) in that the output of the system is a single *optimal* linguistic form that best satisfies a set of conflicting constraints. There are two major differences between OT and the current approach; we briefly discuss these differences before presenting the theory.

In OT, constraints are ranked in a strict-domination hierarchy. This means that, if a constraint α is ranked above a second constraint β, no number of violations assessed by constraint β will be enough to 'justify' a violation of α. In other words, if α and β are the only constraints in the grammar, a form *F* that doesn't violate α will always be preferred to a form that violates α, no matter how many times *F* violates the lower-ranked constraint β. On a first pass, the optimal form is one that violates the highest-ranked constraint fewer times than all other candidates.

The constraints proposed here, in contrast, are weighted. This means that the ranking between α and β is represented as a difference between two real numbers that represent the respective strengths of the two constraints, rather than a categorical domination relation. The constraints attribute a *cost* to any candidate linguistic form that is proportional to the weight of the constraint and the number of times the candidate violates that constraint. The optimal form is one that minimizes the summed cost of violation across all constraints. Given the scenario described above, a candidate form that violates α may emerge as optimal, if in doing so it avoids enough violations of β to result in a lower overall cost. In this respect, the system is more similar to Harmonic Grammar (Legendre *et al.* 1990), a theory that is related to OT but makes use of weighting instead of strict domination.

The second major difference involves gradience. In OT, constraints violations are assessed in a categorical manner, as a discrete number of *marks* assigned to each possible output form, one mark for each instance where a candidate form violates the constraint. The constraints proposed here, in contrast, are *gradiently violable*; they pertain to continuous, non-categorical properties such as the difference between two durations. The cost of violating such a constraint is proportional to the size of the violation. For instance, in the quadratic framework developed here the cost of deviating from a target duration by 90 ms is nine times the cost of deviating from that target by 30 ms, because cost is proportional to the square of the deviance from target. Flemming (2001) introduces this type of framework and uses it to implement several analyses of phonetic and phonological phenomena.

The analysis in this chapter attempts to account for the compression effects discovered in the

production experiment from chapter 2. The tools for accomplishing this analysis will include the weighted constraint formalism, minimal assumptions about the representation of duration and duration targets, and hypothesized differences in the effect of consonants and transitions upon the recoverability of an adjacent vowel. In this manner, the analysis will explain the production asymmetries with reference to independent, falsifiable hypotheses about perception. Some of those hypotheses, in turn, are experimentally tested in the next chapter.

We proceed by considering asymmetries in temporal patterning one at a time, reviewing the possible perceptual explanations for these asymmetries, incorporating those putative perceptual factors into the constraint system, and checking the outputs of the resulting grammar against the data from the production experiment. If the constraint formalism is adequate for characterizing compression patterns, then we expect the outputs of the grammar to be qualitatively similar to the observed data.

Essentially, the objective of this chapter is an existence proof for a grammar that can generate outputs similar to those observed in the production study. The idea is to show that a system of weighted, gradiently-violable constraints on the duration of segments and syllables, when coupled with some assumptions about perceptibility, can derive most of the patterns observed in the experiment reported in chapter 2. Throughout the discussion, we refer to schematic duration patterns measured in arbitrary, abstract units of time. We are concerned almost exclusively with qualitative patterns of relative duration rather than precise quantitative differences. There are several reasons why attempting a more precise simulation would not be fruitful at this stage.

First, milliseconds themselves are a somewhat arbitrary unit from the standpoint of cognition. There is no particular reason to believe that they offer an accurate characterization of perceptual duration; the function from millisecond duration to perceived duration may be linear (highly unlikely), parabolic, or discontinuous in various ways.[1] Duration ratios may offer a more accurate characterization of perceptual duration. Or the perception of duration may be best characterized by some other function we have yet to discover. Attempting to mimic the exact ms values of data from the production experiment will not result in a more principled model than one that captures relative patterns.

Furthermore, the durations reported for the production experiment are in some ways as abstract as the numbers used here. They represent averages over various speakers uttering various tokens at various speech rates. To precisely model minute differences in timing will require some notion of variability on each of these levels; but the *generalizations* across all tokens from all speakers are what particularly interest us here. Those generalizations are taken to be properties of the grammar shared by all speakers. Given that those generalizations are overlaid by several levels of variability, modeling average ms durations would be no less of an abstraction than modeling durations of 10 or 20 arbitrary mental units. It would simply require more fine-tuning of constraint weights, coefficients, and other parameters of the formalism to be discussed here. Put another way, it would greatly complicate the fitting of models without increasing our conceptual understanding of the patterns found in the production study.

[1] The statistical models used in chapter 2 in fact measured duration in standard deviations from a subject's mean value for some set of phonetic forms. This type of measure, which incorporates variability, might also be an interesting way to think about the perception of duration.

Because the weighted-constraint formalism is not entirely familiar, we begin by introducing some of the mechanics of constraint formulation and candidate evaluation in this system in section 3.2. This section also introduces some assumptions and simplifications that will make it easier to find the optimal forms selected by a grammar. The framework is developed to account for patterns of compression in subsequent sections.

3.2 The framework

3.2.1 The constraints

The hypothesis we start with is that the duration of speech units within a syllable is a function of a tradeoff between competing pressures on higher- and lower-level units. In a weighted constraint system, we can construe each of these pressures as a constraint that will assign a fixed *cost* to linguistic forms in proportion to how much they deviate from their target durations. Producing linguistic forms that trade off the two pressures against each other, then, is equivalent to finding forms that minimize the summed cost assessed by the constraints.

The basic intuition behind this system is simple: vowels are shortened as consonants are added in order to keep syllable duration relatively constant.[2] Flemming (2001) implements this idea with two weighted constraints, DURATION-V and DURATION-σ, which assign a cost to linguistic forms that deviate from auditory duration targets for vowel length and syllable length, respectively. This is quite similar to the framework developed here; the evaluation of the duration constraints,

[2] The question of *why* a speaker would want to keep syllable duration relatively constant is a difficult one. It is possible that any tendency toward isochrony in the speech stream, even if imperfect or incomplete, helps a listener parse the speech stream by creating temporal expectations that can be used to guide perception (Quené & Port 2005).

however, will be somewhat more complicated.

We begin with constraints on the duration of higher-level units (syllables, just for concreteness) and lower-level units (segments). As complexity is added to a higher-level unit, these constraints come into conflict. For instance, there's no way to realize 10 segments inside a syllable without either producing a very long syllable or very short segments. The exact tradeoff between avoiding long syllables and avoiding short segments will be determined by the weights associated with the constraints.

The syllable duration constraint can be stated straightforwardly as in (1).

(1) $C_1 = w_1 \bullet (t_\sigma - d_\sigma)^2$

The cost C_1 assessed by constraint 1 is a function of the weight w_1 of the constraint and the difference between the duration target t_σ for a syllable and the actual duration d_σ of the syllable. The difference is squared to eliminate negative numbers and to ensure that cost grows rapidly with increasing deviations from the target, which will result in tradeoffs.

The segment duration constraint is somewhat more complicated, because it requires a special notion of duration for the segment. In fact, the concept 'duration of a segment' is an idealization. While we can pinpoint the division between, for instance, acoustic stop and acoustic vowel with a fair bit of precision, there is still no definitive point in time where the speech signal switches from 'only stop' to 'only vowel'. The problem is more obvious with nasals and especially liquids

adjacent to vowels. To incorporate this observation into the model, we allow the duration target for a segment to be partially satisfied by information contained outside the 'segment proper'. The recoverability of a segment is related to its duration *and* cues to its presence or its features that are contained in other parts of the signal. In this approach, what we refer to as the 'duration' of a segment is really more like a segment's recoverability over time.

For instance, the recoverability of a vowel will be directly proportional to the duration of its acoustic steady state times some constant i, plus the duration of the transition to an adjacent segment times some constant $j < i$, plus the duration of the adjacent segment times some coefficient $k < i$, where j and k vary across different manner features. The vowel-recoverability coefficients j and k represent the relative amount of information about a vowel contained in an adjacent transition and adjacent segment, respectively. They represent something like the 'vowel transparency' of those intervals. By hypothesis, liquids have higher vowel transparency than obstruents and possibly nasals; these predictions are based on the production data presented in the preceding chapter. We constrain k and j to be lower than i, at least for vowels, because we assume that internal cues (consisting largely of formant frequencies) contain more information about a vowel than external cues in adjacent portions of the speech stream.

Conceptually, i, j, and k should be construed as coefficients that correlate with the amount of 'vowel information' contained in any given stretch of an utterance. Vowel information itself will not be fully explained here; assume for the time-being that it is more or less directly reflected in subjects' ability to discriminate vowel contrasts at any point in the speech stream. Given these hypotheses, the segment duration constraint can be stated as in (2).

(2) $C_2 = w_2 \bullet (t_s - (id_s + jd_t + kd_a))^2$

The cost C_2 assessed by this constraint is a function of its weight w_2 and the difference between the target duration and actual duration. However, 'actual duration' here is something more like recoverability: it is computed as the sum of some coefficient *i* multiplied by the duration of the segment proper d_s, some coefficient *j* multiplied by the duration of the adjacent transition d_t, and some coefficient *k* multiplied by the duration of the adjacent segment d_a. This constraint will apply in turn to each segment in a candidate linguistic form.

For the analyses in this chapter, we will generally use values of *j* and *k* between 0 and 1, with *i* implicitly set to 1. There is no particular hypothesis behind this; it merely seems like an intuitive scale to use. If a vowel has an internal recoverability coefficient of 1 (proportional to the amount of information the vowel contains about itself), an adjacent liquid might have a *k* value of 0.6, reflecting relatively high vowel transparency, while a stop might have a value of 0.1. Similarly, we might assign a *j* value of 0.6 to modally-voiced formant transitions into or out of a vowel, and assign a value of 0.2 to formant transitions overlaid by nasalization. These numbers are meant only to suggest relative patterns.

Given the constraints in (1-2) and a set of values for the variables contained within them, we can assign a cost to any candidate linguistic form. In order to find out what types of linguistic forms are predicted to actually surface, we need to find candidates that minimize the summed cost assessed by constraints. Given a set of values for the parameters that the constraints make reference to, only one form will emerge as optimal. In the next section we examine several ways

of finding that form.

We represent the differences between various types of syllable by assigning them different parameter settings; for example, we represent the difference between *la* and *da* as a difference in the vowel-recoverability coefficient *k*, as mentioned above. Using a variety of different sets of parameter values as input to the model, with one set of values for each type of syllable, we can then observe the array of surface duration patterns predicted by the constraint system. These will be compared to the experimental data.

3.2.2 Finding a winner

The constraints stated above will assign the following cost to any syllable σ consisting of consonant *x* and vowel *y*:

(3) $w_1 \bullet ((d_x + d_t + d_y) - t_\sigma)^2 + w_2 \bullet (nd_y + md_t + ld_x) - t_x)^2 + w_2 \bullet ((kd_x + jd_t + id_y) - t_y)^2$

where variables *i-n* are recoverability coefficients for various portions of the speech stream.

Expression (3) sums the violations from the syllable constraint (first quadratic term), the segment constraint applied to the consonant *x* (second term), and the segment constraint applied to the vowel *y* (third term). For most of the analyses in this chapter, we will assume that the same segment constraint applies to consonants and vowels, and therefore has the same weight. This is

because a system with only one segment constraint is complex enough to capture most of the data, and there is no reason to add in more complexity at this point.

For the purposes of the current analysis, we simplify the cost function in several ways, in order to make the optimization problem more tractable. Because we're mainly interested in the effect of context on *vowel* duration, we'll remove the recoverability coefficient terms from the consonant duration target, which reduces the number of free parameters in the model. Also to simplify, we'll assume that the recoverability coefficient *i* for a segment itself is 1. Of course, for segments like stops that have more external cues than internal ones, this may not be realistic.

With these simplifications, the cost function that we need to minimize in order to find an optimal phonetic form is as shown in (4).

(4) $$\text{Cost} = w_1 \bullet ((d_x + d_t + d_y) - t_o)^2 + w_2 \bullet (d_x - t_x)^2 + w_2 \bullet ((kd_x + jd_t + d_y) - t_y)^2$$

The approach is to assign values for all parameters except the segment durations d_x and d_y as input, then determine which values for those durations will minimize the cost function. This means that we are assuming values for everything except segment steady-state duration (and syllable duration, which depends on it), and treating phonetic forms with various steady-state durations as candidate realizations. Any pair of consonant and vowel durations is a possible candidate; only the one that incurs the smallest cost is the optimal candidate.

In what follows, parameters that are given values as an input are referred to as *constants*. Again, this refers to all parameters except the segmental and syllabic durations. Once we've picked values for the constants in this expression (constraint weights, target durations, etc.), we can examine how the cost function changes across different values for the actual duration of *x* and *y*.

For the moment, we'll set the constants as follows: w_1 (syllable constraint) = 1; w_2 (segment constraint) = 2; t_σ = 30 arbitrary duration units (ADU); t_x = 15 ADU; t_y = 25 ADU; d_t (duration of the transition between *x* and *y*) = 4 ADU; k = 0.2; j = 0.4. The only crucial assumption embedded in these numbers is that the sum of the duration targets for segments is greater than the duration target for syllables; without this property, of course, the system will not predict compression.

Note that we assume a constant transition duration; this corresponds to the hypothesis that transitions are essentially interpolation between targets and are not under active control by the grammar. That assumption could be changed if need be, but we'll begin by manipulating the fewest number of parameters that are necessary to account for the data.

We now examine four candidate phonetic realizations. Candidate A has relatively much vowel shortening, in order to better satisfy the other constraints; candidate B shortens the consonant a lot to better satisfy the other constraints; candidate C lengthens the syllable substantially to accomodate both of the segments; and candidate D shortens both segments and also lengthens the syllable, all in moderation.

Candidate	C duration	V duration	Con 2 V	Con 2 C	Con 1	Total Cost
A	14	16	42.32	2	16	60.32
B	10	20	3.92	50	16	69.92
C	14	20	0.72	2	64	66.72
D	12	18	18	18	16	52

Table 3.1. *Cost assessed to four hypothetical candidate phonetic realizations of a CV syllable. Columns contain candidate name, realized duration of C and V, cost assessed to V and C by constraint 2, cost assessed to the syllable by constraint 1, and total cost assessed to the candidate. Parameter settings are as indicated in the text above.*

As can be seen in table 4.1, this system disfavors candidates that egregiously violate any of the constraints, as in A-C, and favors candidates that 'compromise' by violating each constraint in moderation, as in D. For a more complete picture, we can examine cost as a function of d_x and d_y in three dimensions:

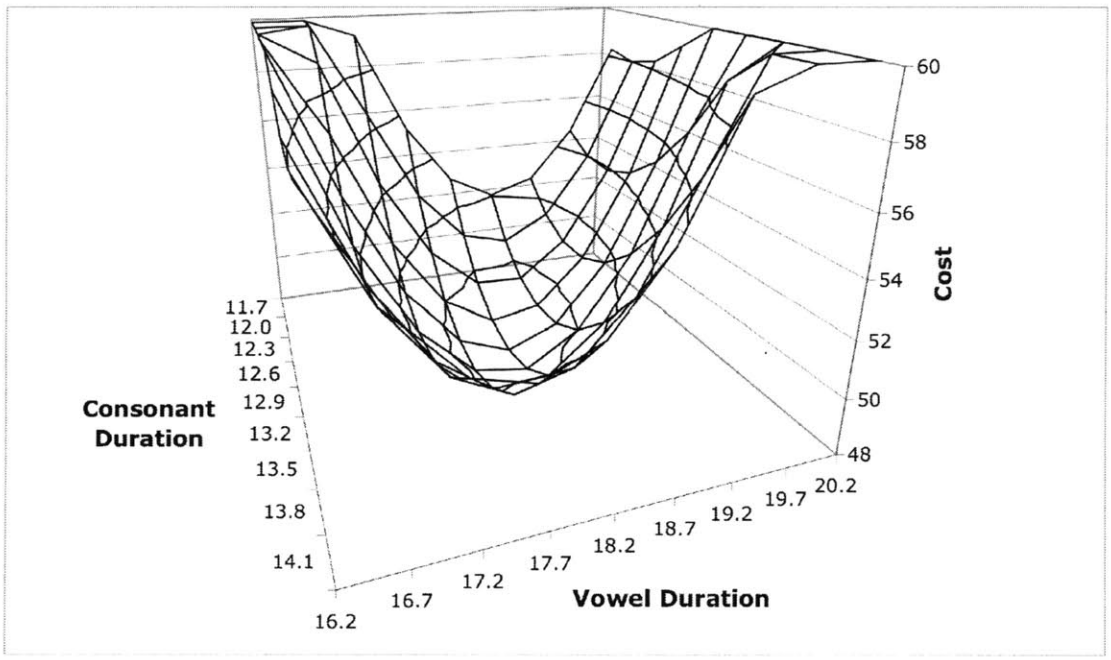

Figure 3.1. *Cost as a function of consonant and vowel duration for a CV syllable, with constants specified above.*

Figure 3.1 shows cost as a function of segment steady-state durations; our cost function maps every point in the two-dimensional 'durational space' represented by the floor of the graph to a cost associated with the candidate that is realized with those segment durations. Cost is represented on the vertical axis. The cost function takes the form of a bowl. The point of lowest cost, corresponding to the optimal phonetic form, is the point at the bottom of the bowl. It may be easier to visualize the problem as a relief map viewed from directly above the bowl; this graph is similar to a topographical map of elevation.

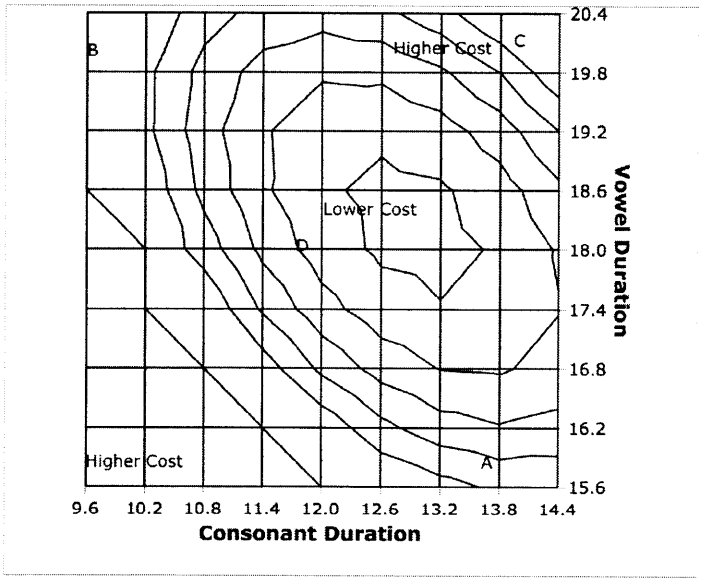

Figure 3.2. *Cost as a function of consonant and vowel duration for a CV syllable, with constants specified above. This is figure 3.1 viewed from directly above as a relief map. The area at the center is lowest, with cost increasing in concentric rings outward from the center. The letters A-D show the approximate locations of the four candidates from table 4.1.*

To determine the predictions of the grammar, we need to determine where the bottom of the bowl is located in durational space. We can arrive at an approximate solution by examining these

graphs (or the tables that they're derived from) and identifying the lowest point. Here, it is somewhere in the area of (12.9, 18.2). This method of identifying optima, however, is unreliable and inefficient. The exact values that we find will depend on the granularity of our chart. Regenerating a new table or graph every time we change a parameter setting and searching for the low point by hand is also time-consuming and prone to errors. Fortunately, we can also identify optima analytically by examining the cost function itself.

To find the optimal durations for segments *x* and *y* with regard to these constraints, we need to find the bottom of the bowl. Abstractly, that bottom point can be defined as the point where the bowl stops sloping along the consonant axis and the vowel axis; in other words, it is the point where the instantaneous slope along both axes is equal to zero. To state the instantaneous slope of the cost function at any point along each of the two axes, we'll need to consider the partial derivative of (4) for each variable. These will tell us how the cost assessed by constraints changes as we change one variable, holding the other constant. Quadratic terms were chosen for the constraints in part because they have linear derivatives and are relatively easy to differentiate: the derivative of x^2 is 2x. The two partial derivatives are shown in (5-6).

(5) $f'_x(d_y) = 2w_1 \bullet ((d_x + d_t + d_y) - t_\sigma) + 2w_2 \bullet ((kd_x + jd_t + d_y) - t_y)$

(6) $f'_y(d_x) = 2w_1 \bullet ((d_x + d_t + d_y) - t_\sigma) + 2w_2 \bullet (d_x - t_x) + 2w_2k \bullet ((kd_x + jd_t + d_y) - t_y)$

These expressions give us the instantaneous slope along the two segmental duration axes at any point in durational space, as a function of the model parameters. We obtain them by differentiating the full cost function for one variable at a time (d_y in (5) and d_x in (6)) while

treating the other variables as constants. We're looking for a point where both of the partial derivatives are equal to zero. This is a point where the cost function is instantaneously flat or parallel to the floor; in other words, the bottom of the bowl. First we set one derivative to zero:

$$(7) \qquad 0 = 2w_1 \bullet ((d_x + d_t + d_y) - t_\sigma) + 2w_2 \bullet ((kd_x + jd_t + d_y) - t_y)$$

We then solve for the duration of y in terms of x and the constants:

$$(8) \qquad d_y = \frac{w_1 \bullet ((d_x + d_t) - t_\sigma) + w_2 \bullet ((kd_x + jd_t) - t_y)}{-(w_1 + w_2)}$$

Expression (8) shows that the duration of the vowel varies inversely with durations of the consonant and transition; this is exactly the hypothesis that our constraint system was intended to implement. More precisely, the duration of the vowel is proportional to the amount of syllable target not filled by the consonant and transition, $d_x + d_t - t_\sigma$; and the amount of vowel target not filled by the coefficient-adjusted duration of the consonant and transition, $kd_x + jd_t - t_y$. Substituting the expression in (8) back into (6) will allow us to solve for the duration of x solely in terms of the constants. This means that, given any set of parameter values, we'll be able to determine the optimal durations for consonant and vowel exactly, without resorting to the trial and error method used above.

$$(9) \qquad d_x = \frac{w_2 \bullet (1 - k) \bullet (t_\sigma - d_t - (t_y - jd_t)) + t_x \bullet (w_1 + w_2)}{w_2 \bullet (k^2 - 2k + 2) + w_1}$$

Expression (9) shows that the optimal duration for the consonant x is a rather complicated function of the values associated with targets, constraint weights, and recoverability coefficients. It is positively correlated with its own target duration, as we would expect. It is also positively correlated with the quantity of syllable target not accounted for by transition duration, $t_\sigma - d_t$, reflecting the fact that there will be more room for the consonant when the transition is shorter. It is negatively correlated with the quantity of vowel target not accounted for by the transition, $t_y - jd_t$; this is because the vowel will need to be longer when the transition is less informative, leaving less room for the consonant. When we plug the values we used above into expressions (8) and (9), it returns values of about 12.93 for d_x and 18.23 for d_y. This is consistent with what we concluded from the graphs in figures 3.1 and 3.2.

With the analytical solution in place, we no longer need to examine cost tables or charts to determine output forms. In what follows, the optimal values for outputs will simply be presented as predictions of the model.

3.3 Simplex CS phenomena

At this point, we can already begin to analyze some of the phenomena from the production experiment. Because so far we only have a cost function in place for two segments, we can't address complexity effects yet. Some relevant patterns, however, obtain between items with the same number of segments.

Recall that the duration of the vowel in VC sequences depended in part on the quality of the transition from the vowel in *the* into the initial vowel of the target word. When that transition came in the form of a glottal stop (henceforth *closure*), the following vowel was much longer than when the transition consisted of creaky or modal formant transitions (henceforth *no closure*). In the current framework, we characterize these two realizations as containing no onset consonant *x* but only a transition. The difference between the two realizations lies in the vowel transparency of the onset transition, as indicated by the vowel-recoverability coefficient *j*. By hypothesis, formant transitions contain more information about a following vowel than silence followed by glottal release does; as such, we assign a higher *j* value to tokens with the former realization.

Comparing the outputs of the model for VC stimuli with *j* set at 0.1 and 0.8, we predict that items with higher *j* should have shorter vowels. This means that the steady-state modally-voiced vowel should be shorter following a glide or formant transition realization than following a realization with closure. The predictions match the data rather closely. We retain the parameters (except *j*) from the previous section here: $w_1 = 1$; $w_2 = 2$; $t_\sigma = 30$; $t_x = 15$; $t_y = 25$; $d_t = 4$; $k = 0.2$.

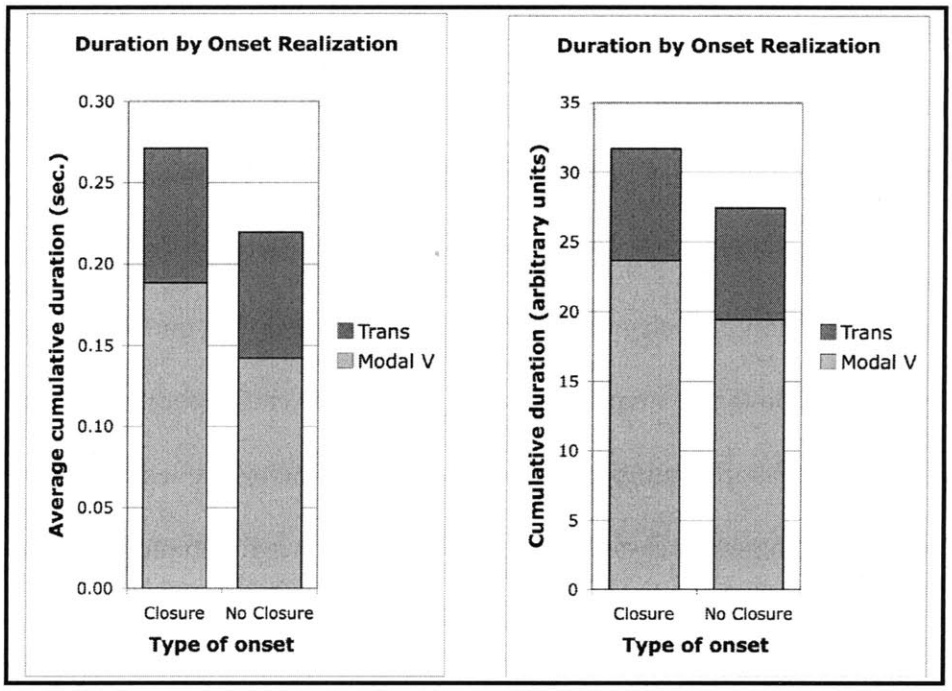

Figure 3.3. *Data from the production experiment (left) compared to the predictions of the constraint system (right) concerning the variable realization of VC items. 'Trans' indicates the duration of all portions of the syllable up to the onset of modal voicing with steady formants.*

Note that this analysis assumes a somewhat arbitrary two-way distinction between stop and glide realizations. In reality, the transitions observed in this context fall along a continuum, as discussed in section 2.3.1. We could incorporate this observation by assigning a continuum of *j* values to different realizations, which would correlate with the duration and auditory clarity of formant structure contained therein; the general prediction would be that realizations with higher vowel transparency are followed by shorter steady-state vowels. There are several reasons why we haven't made such fine-grained distinctions here. The various realizations are not balanced across subjects, vowels, or presumably speech rates; some of the realizations are represented by very few tokens; consequently, we probably don't have enough data for statistical testing of all the observed patterns; and the binary distinction used here is enough to illustrate the qualitative

predictions of the model. A follow-up experiment might elicit a larger number of VC items and come to more statistically solid generalizations about the relationship between onset realizations and vowel duration, but the current study was not designed for this purpose.

In figure 3.3 and throughout this chapter, we are more concerned with qualitative patterns of greater and lesser duration than with precise quantities. We've chosen values for the duration of transitions and vowels that are similar in their proportions to the ms values from the production experiment, to make such graphs easier to compare. With further fine-tuning of target durations, constraint weights, and recoverability coefficients, we could come trivially close to predicting the actual values observed in the experiment, but this wouldn't gain us anything for the reasons discussed in section 3.1. What is important in the above graphs is that we observe shorter steady-state vowels following one type of phonetic realization than following another, and that the model, when supplemented with a perceptual hypothesis about the two realizations, predicts that this pattern should hold.

A second pattern concerns the vowel-recoverability coefficient *k* associated with the vowel transparency of an adjacent consonant steady-state. We can check the model's predictions for the difference between consonants with a high value for *k* and a low one. Again, we predict a shorter steady-state vowel adjacent to the segment with a higher coefficient. By hypothesis, liquid steady-states contain more vowel information than stops. The graph below compares consonants with *k* values of 0.1 and 0.4 to actual data from onset /p^h/ and liquids, respectively. The qualitative match, again, is rather good. We assume for the purposes of the simulation that a vowel adjacent to no onset consonant will simply be realized with its target duration.

Alternatively, we could say that it will be realized with a duration that is a weighted average of the segment and syllable targets; this makes no qualitative difference in what follows.

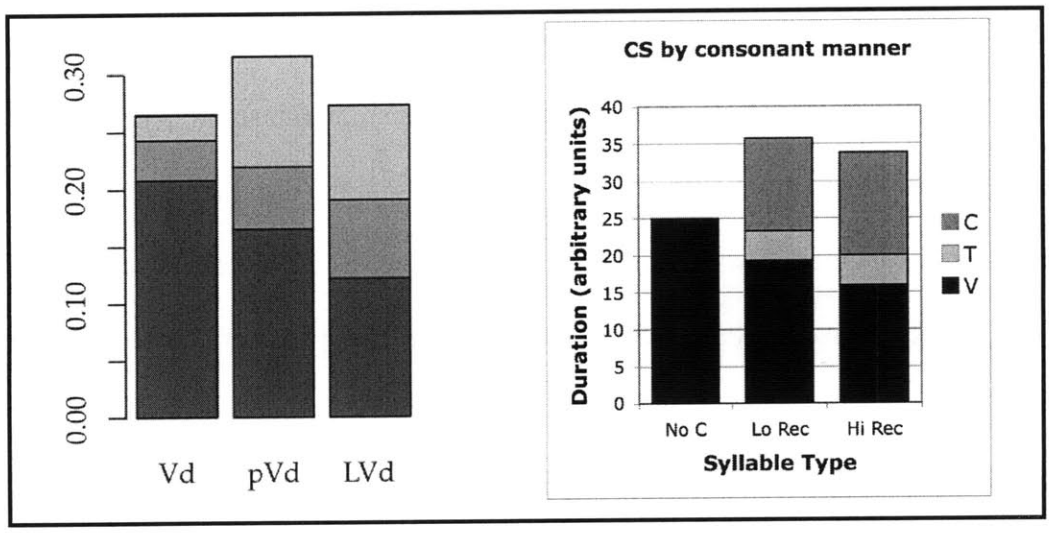

Figure 3.4. *Data from the production experiment (left) and model predictions (right) for consonant manners with high (rightmost bars) and low (center bars) vowel-recoverability coefficients. For production data, durations are in seconds. The upper bars for vowel-initial items represent closure and transition durations, in realizations where these categories are applicable.*

The qualitative predictions of the model for simplex CS effects match experimental data fairly well. In order to generate predictions for complex CS, a few more analytical steps are necessary. These are outlined in the next section.

3.4 Complex CS phenomena

3.4.1 Scaling up the optimization routine

The same optimization procedure outlined above will apply to cases with three, four, and up to *n* segments. However, every time we add another segment there will be more partial derivatives to solve and each one will be more complicated. For this reason, the discussion here will be limited to three segments. Making the same simplifying assumptions as we did earlier, the cost function for a CCV syllable consisting of string *xyz* will be as in (10). We simplify further by assuming that the adjacent consonant contains vowel information but the non-adjacent one does not. This assumption is plausibly incorrect, but the simplified function will suffice to derive predictions.

(10) $w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma)^2 + w_2 \bullet (d_x - t_x)^2 + w_2 \bullet (d_y - t_y)^2 + w_2 \bullet (kd_y + jd_t + d_z - t_z)^2$

In expression (10), the first quadratic term concerns syllable duration; the other three terms concern outer consonant *x*, inner consonant *y*, and vowel *z*, respectively. The partial derivatives of this expression are shown in (11).

(11) $f'_{y,z}(x) = 2w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma) + 2w_2 \bullet (d_x - t_x)$

$f'_{x,z}(y) = 2w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma) + 2w_2 \bullet (d_y - t_y) + 2w_2k \bullet (kd_y + jd_t + d_z - t_z)$

$f'_{x,y}(z) = 2w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma) + 2w_2 \bullet (kd_y + jd_t + d_z - t_z)$

Each of these expressions treats one segment duration as a constant and the other two as variables. Conceptually, they characterize how the cost changes as a function of two segment

durations, while holding the third constant. Following the same procedure as before, we solve out for d_z in terms of the constants and the durations of the other two segments. The answer is:

(12)

$$\frac{d_y \bullet (w_2 \bullet ((k+1) \bullet k + 1) + 2w_1) + 2w_1 d_x + d_t \bullet (w_2 j \bullet (k + 1) + 2w_1) - w_2 t_y - w_2 t_z \bullet (k + 1) - 2w_1 t_\sigma}{w_2 \bullet (k+1) + 2w_1}$$

It should already be clear that the equations and the algebra involved in this analysis are a fair bit more complicated than the previous example. The rest of the system of equations is shown in appendix 3A. The subsequent steps involve substituting the expression in (12) back into one of the derivatives in (11) to solve for *y* in terms of *x*, then repeating this solve-and-substitute process until one of the partial derivatives can be expressed solely in terms of a single variable. Setting that derivative equal to zero allows us to solve for that variable solely in terms of the constants, and the remaining variables can then be bootstrapped from the first one.

3.4.2 Incremental CS

The model as it is currently stated will always predict CS when an extra segment is added into the syllable. This is because avoiding egregious violations of the syllable target will always justify some amount of segment shortening; changing the weights of the constraints or the vowel-recoverability coefficients will only affect *how much* shortening is observed: higher coefficients for adjacent intervals result in more shortening.

We retain the parameter setting from the preceding sections: $w_1 = 1$; $w_2 = 2$; $t_\sigma = 30$; $t_x = 15$; $t_y = 25$; $d_t = 4$; $k = 0.2$; $j = 0.4$. We assume as well that the two consonants in a cluster have the same duration target values. These settings will derive the following phonetic realizations for a V - C_1V - C_2C_1V triplet:

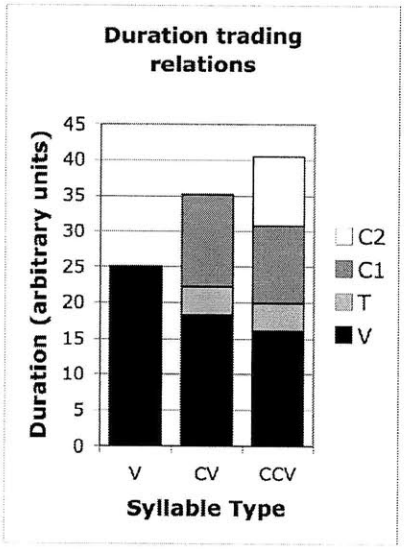

Figure 3.5. *Incremental CS as predicted by the constraint system.*

This analysis, then, accounts for cases where incremental CS is observed. If the vowel-recoverability coefficient *k* or the weight of the syllable constraint were higher, we would predict more incremental CS. As currently formulated, however, the analysis can't account for the cases where *no* incremental CS is observed. To explain those cases, we need to import an assumption from the Klatt (1979) duration model. In that model, linguistic objects are associated not only with an inherent target duration, but a minimum duration past which they may not shorten. This corresponds to the idea that a segment must have some minimum duration if its presence and/or quality is to be detected. If this standard idea is incorporated into the model, we predict that for

some pairs of comparable items CS will not be observed, because the vowel will be unable to shorten any more without crossing the minimum duration threshold. In what follows, this will be referred to as a *floor effect*. Note that the effects here correspond to recoverability floors rather than duration floors, because they can be partially satisfied by external cues.

One way to incorporate floor effects into the model is to simply recast the vowel-duration constraint as a discontinuous function that penalizes durations above the minimum in its normal fashion, but assigns maximum cost to any durations below the minimum. Consider how this affects candiate evaluation if we set k to 2, w_1 to 5, the floor threshold to 20, and keep all of the other values the same. The cost function now looks as shown in the two figures below:

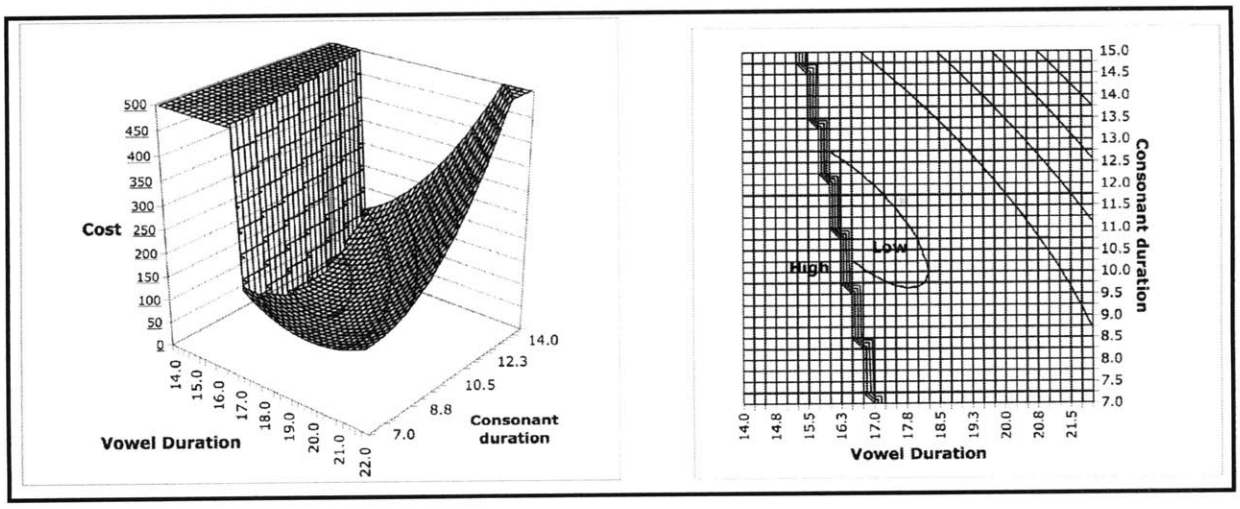

Figure 3.6. *Cost function for a CV syllable with a vowel floor, in three dimensions (left) and viewed as a relief map from above (right).*

The minimum recoverability requirement has the effect of throwing up a 'wall' in durational space. If the vowel target only depended on the vowel proper, this wall would be perpendicular to the vowel axis; because of the vowel-recoverability coefficient, however, it is oriented at a

slight diagonal in durational space. For this particular set of values, the wall hasn't blocked off the bottom of the cost function's bowl; the optimum here is still the same as it was before. Consider what happens, however, when we add another consonant in:

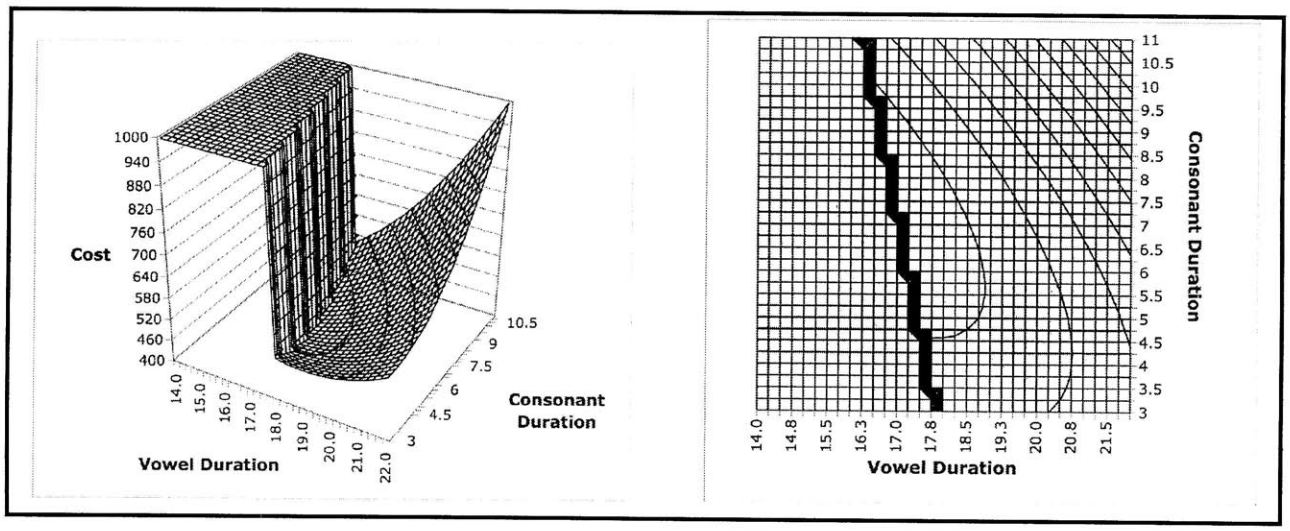

Figure 3.7. *Cost function for a CVCC syllable with a vowel floor, in three dimensions (left) and as a relief map from above (right). C2 duration held constant at 5.5 units.*

The wall has now 'cut off' what would have been the bottom of the bowl in a model without floor effects. In other words, the minimum vowel duration keeps the grammar from selecting a compressed form that would otherwise have been optimal; the grammar in some sense doesn't have access to duration patterns that lie on the other side of that wall. The vowel-recoverability coefficients, because they affect how much of the minimum duration needs to be filled by steady-state vowel duration, correspondingly affect the location and orientation of the wall in durational space. Adjusting the vowel-recoverability coefficient k has the effect of changing both the absolute position and angle of the wall relative to the vowel duration axis. This, in turn, affects which part of the original cost function's bowl is 'cut off from the grammar'. The result is

that consonants with different values of *k* induce different amounts of incremental CS: the higher the *k* value, the more incremental CS is observed.

Shown below are model predictions and actual data for liquids and obstruents in CVC and CVCC syllables. The obstruents show no incremental CS, while the liquids do. The crucial conditions for deriving a difference in the presence of incremental CS between different manners of consonant are as follows: the consonants or their transitions differ in their vowel-recoverability coefficients; and the CVC form for the consonant with lower *k* is sufficiently close to the vowel floor to preclude further vowel shortening.

The values for this particular simulation are $w_1 = 1000$; $w_2 = 10$; $t_\sigma = 35$; $t_{x,y} = 11$; $t_y = 25$; $d_t = 4$; $j = 0.4$; vowel floor is 23.1; consonant floor is 7. The vowel-recoverability coefficient *k* is set at 0.1 for obstruents and 0.6 for liquids. These values were arrived at by attempting to minimally modify the values used in earlier sections; the floor values are fine-tuned to derive zero shortening in the low-coefficient case.

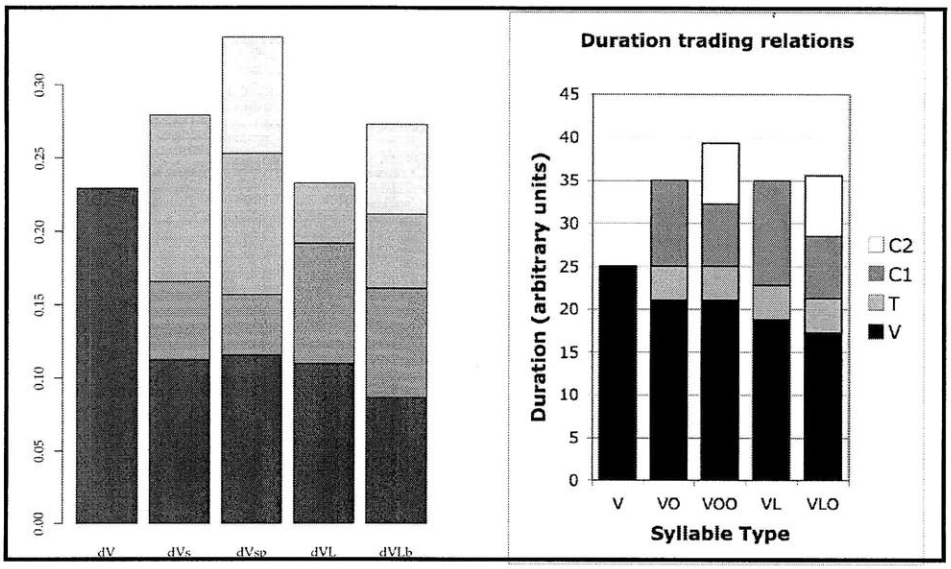

Figure 3.8. *Production data (left) and model predictions (right) for obstruents and liquids in CVC and CVCC syllables. Durations for production data are in seconds. In predicted data, 'V' = vowel; 'O' = obstruent; 'L' = liquid.*

Note that in the production data shown above, the transitions from vowel to liquid are clearly longer than those from vowel to fricative. This observation is confounded, however, by the different measurement criteria used in the two contexts: boundaries generally track F3 for /r/, and different combinations of F1 and F2 in other contexts. If the difference in transition duration is reflecting a real perceptual property rather than measurement differences, then the model's predicted asymmetry in CS shown here would hold *a fortiori*, because the prolonged transitions in the liquid case would contain more vowel information.

As it stands, the analysis of floor effects relies on what are essentially separate constraints for a segment's target duration and its minimum duration. This is not entirely desirable, because it attributes to the grammar two independent ways of penalizing the same property, namely making a segment too short. It would be possible to unify these two constraints by stating them as an

asymptotic hyperbola function. Such a constraint penalizes segments in proportion to the *reciprocal* of how far they deviate from the minimum. As a segment gets shorter and shorter, the assessed cost 'ramps up' slowly. At a certain point near the minimum, cost spikes suddenly and approaches infinity. This is shown in two views below; the logarithmic scale makes it easier to see the slow change in cost at values well above the target (which is now the same as the minimum), while the normal scale makes it easier to see the sudden spike in cost near the target.

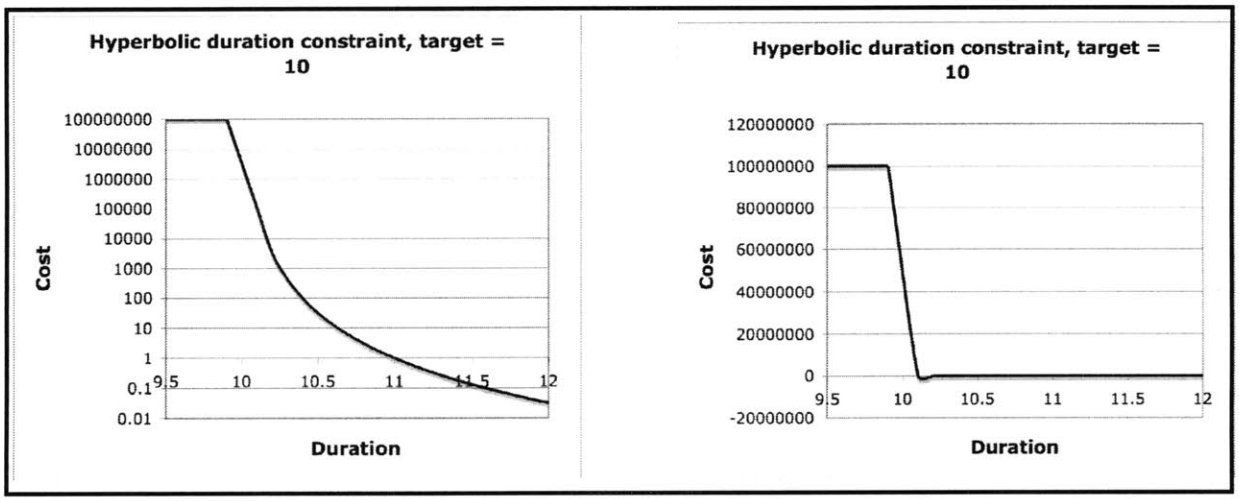

Figure 3.9. *Hyperbola function for a duration constraint, with a target of 10. Cost is shown on a logarithmic scale on the left, on a normal scale on the right. This particular constraint assesses cost equal to the reciprocal of the segment's deviation from target raised to the fifth power. Negative costs, corresponding to segments shorter than the minimum duration, are declared to be equal to* 10^8.

This type of constraint has several desirable conceptual properties. It reflects a 'hard' floor past which no segment can shorten (as the cost would be infinite), and predicts that added duration above the floor will reduce the cost of candidates, but will generate diminishing returns the more the segment is lengthened. Shown below is a cost function for CVC syllables and duration

predictions of a model with hyperbola constraints for obstruents and liquids. It turns out that this system can only simulate the experimental data with different weights for the vowel-duration constraint and the consonant-duration constraint. Distinguishing between the two types of segment constraint might be necessary in any system in order to make precise quantitative predictions about the magnitude of effects. Values are as follows: $w_1 = 5000$; $w_{2C} = 1$; $w_{2V} = 5000$; $t_\sigma = 33$; $t_{x,y} = 10$; $t_y = 20$; $d_t = 4$; $j = 0.4$.

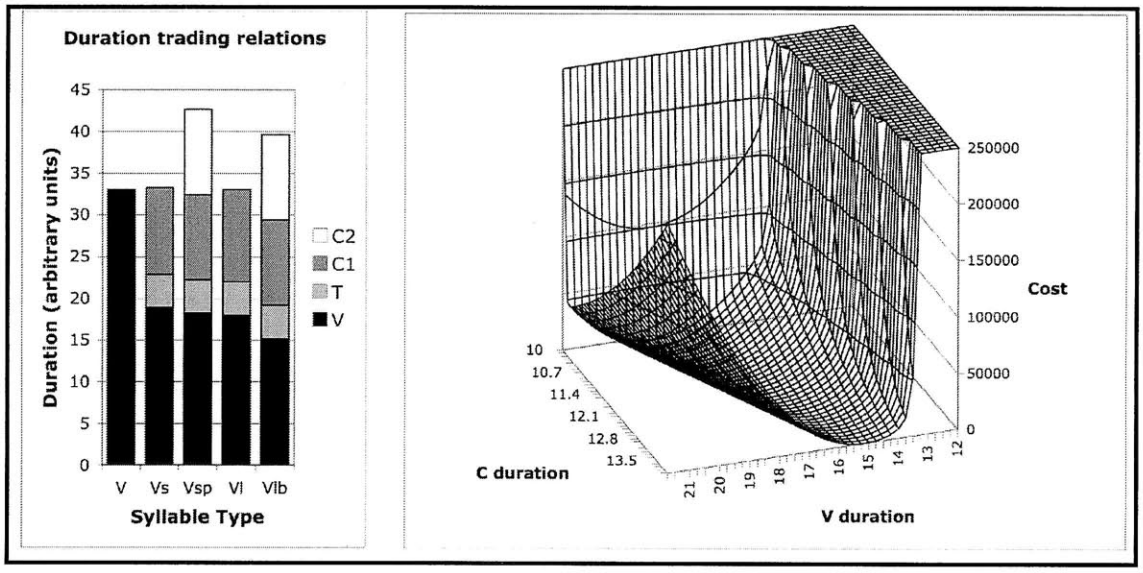

Figure 3.10. *Duration predictions (left) and CVC cost function (right) for a system with the hyperbola segment-duration constraints introduced above.*

Although these constraints are formally and conceptually more elegant than the 'brute force' option of stipulating a floor value as in the earlier system, they also entail several practical difficulties. First and foremost, the cost functions generated by such a system are orders of magnitude more difficult to analyze than the earlier system. This is because expressions with fractions have more complicated partial derivatives, in turn requiring more complicated algebra to solve out the resultant system of equations. We can reach approximate solutions for particular

values of constants by analyzing charts as in the sections above, but, as mentioned in those sections, this is inefficient and unreliable. It also turns out to be rather difficult to fine-tune the balance between how fast the cost assessed by various constraints ramps up as candidate forms approach the minimum duration. This is why the analysis required separate constraint weights for consonant and vowel target constraints. For these reasons, we will continue to use the parabola constraints that were introduced first.

3.5 Other asymmetries

Several other asymmetries were observed in the production data: liquids induce more incremental CS in coda than in onset position; nasals induce incremental CS in onset but not coda position; and vowel steady-states are shorter following /sp/ clusters than following /p^h/. For the first two cases, this may be explained if the vowel-transparency of the consonant steady state, represented here by coefficient *k*, differs systematically between onset and coda position. In that case, the predictions would be much the same as the asymmetries in incremental CS addressed in the preceding section: more CS for higher values of *k*. This might be plausible for liquids, due to the articulatory differences between initial and final liquids discussed in section 2.4.1: liquids involve weaker tongue-tip constrictions in final position; plausibly, this could allow the vowel to influence the acoustic signal more. If this is the correct explanation, however, it would predict the same asymmetry for nasals, because they display the same tongue-tip constriction asymmetries. Instead, we found that they induce incremental CS in onset but *none* in coda position.

Because explanation of these asymmetries in terms of the perceptual properties of consonant steady-states are problematic, we should also entertain the possibility that they are due to asymmetries in transitions between contexts. As mentioned in section 2.4, liquid-vowel transitions in onset position are overlaid by fewer liquid gestures than comparable transitions in coda position. This is true as well for nasals, but the relevant property is that nasalization intrudes more upon a preceding vowel than a following one. So the perceptual properties of transitions may prove a more useful explanation of the observed production asymmetries. In this section, we demonstrate that perceptual differences in the vowel-transparency of transitions could in principle explain those effects.

For liquids, there was an asymmetry with regard to syllable position: incremental CS was greater in coda position than in onset position. The hypothesis put forward to explain this asymmetry was that transitions between vowel and liquid contain more information about the vowel in coda position than in onset position. In the current model, this hypothesis is represented as a difference in the transition vowel-recoverability coefficient *j* between the two contexts. Shown below are production data and model predictions for this onset-coda asymmetry. Parameters for the model are the same as above except for *j*: $w_1 = 1000$; $w_2 = 10$; $t_\sigma = 35$; $t_{x,y} = 11$; $t_y = 25$; $d_t = 4$; $k = 0.6$; vowel floor is 23.1; consonant floor is 7. Values for *j* are set to 0.6 in onset position and 0.8 in coda position.

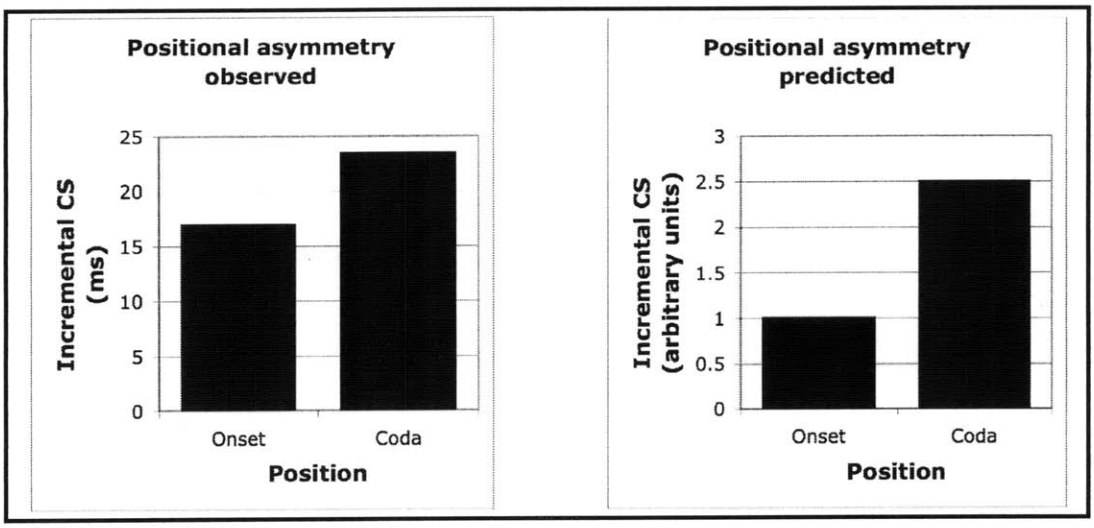

Figure 3.11. *Production data (left) and model predictions (right) for liquid-driven incremental vowel shortening in onset and coda position.*

This simulation shows that a difference in the amount of vowel information contained in pre- as opposed to post-vocalic transitions could be used to explain the observed differences in vowel duration between the two contexts. The production-data graphs also show that the steady state of the *consonant* tends to be longer in CVCC items than in CVC items; recall, however, that this effect is confounded with measurement differences due to the following segment, and is common to all liquid and nasal items.

Another syllable-position asymmetry was observed for nasals. There is incremental CS for clusters involving nasals in onset position, but not in coda position. It was hypothesized that this may be due to a difference in the informativity of transitions adjacent to nasals between onset and coda position.

These facts can be accounted for by attributing a higher vowel-recoverability coefficient *j* to n-V transitions than to V-n transitions. Shown below are production data and model predictions. Parameters for the model are $w_1 = 1000$; $w_2 = 10$; $t_\sigma = 33$; $t_{x,y} = 11$; $t_y = 25$; $d_t = 4$; $k = 0.2$; vowel floor is 21.7; consonant floor is 7. Values for *j* are set to 0.8 in onset position and 0.1 in coda position.

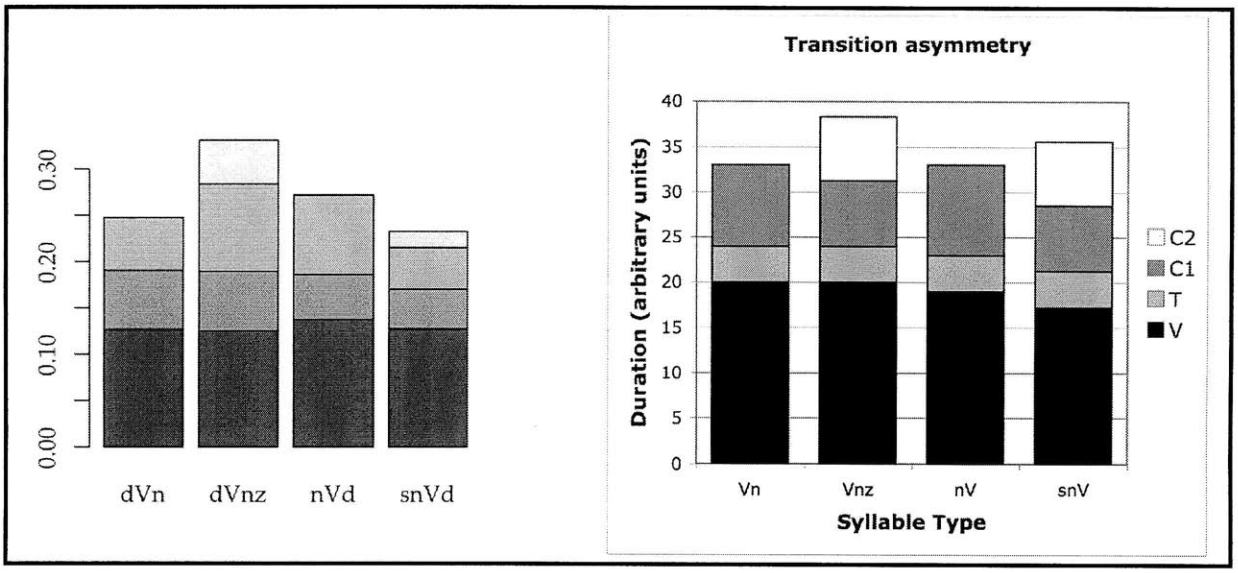

Figure 3.12. *Production data (left two graphs) and model predictions (right) for nasals in onset and coda position. Durations for production data are in seconds.*

One final effect concerns items with obstruents in onset position. We observed a tendency for singleton voiceless stops to be associated with longer steady-state vowels than /sp/ clusters. This is not a straightforward incremental CS effect, because the two sets of items differ in more than just the number of segments contained in the syllable. They also differ in the realization of those segments: the singleton stops involve aspiration, while the clusters instead contain modally-voiced formant transitions into the vowel.

This could plausibly reflect the presence of the /s/ in clusters, which might contain some vowel information even though it is not directly adjacent. The hypothesis to be explored here, however, is that vowel steady-states are shorter adjacent to formant transitions because formant transitions contain more vowel information than aspiration does.

If we treat aspiration as a transition between stop and vowel, we would assign it a lower vowel-recoverability coefficient than the formant transitions in the /sp/ cluster. This corresponds to the hypothesis above that modally-voiced formant transitions contain more vowel information than aspiration. The grammar as currently formulated could predict the observed pattern based only on a difference between the informativeness of transitions; this is shown below. In terms of the model, this contrast is formally identical to that between a singleton consonant with a low value for the transition vowel-recoverability coefficient *j* and a cluster with a higher *j* value. As such, it is equivalent to the contrast between Vn and snV in the previous simulation; parameters are identical to those used above.

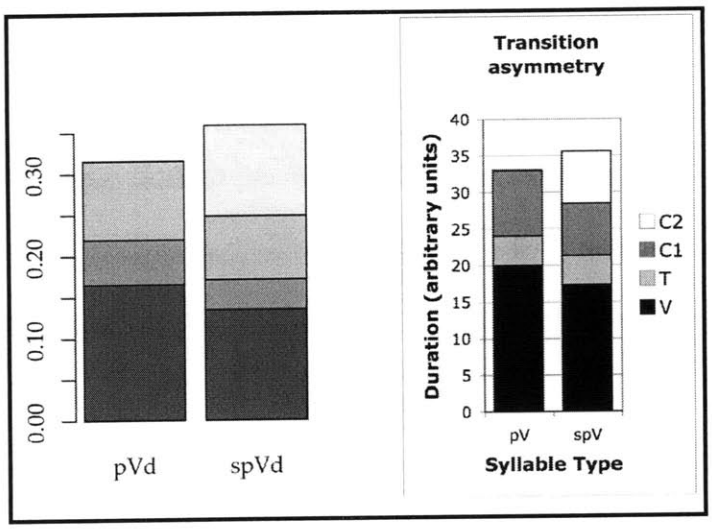

Figure 3.13. *Production data (left) and model predictions (right) for aspirated stops and /sp/ clusters. Durations for production data are in seconds.*

An additional difference is visible in the production data in figure 3.13: aspiration is generally longer than formant transitions. Even when such a difference is input into the model (in the form of different fixed transition durations), it can still predict that the vowel adjacent to the /sp/ cluster will be shorter. This result holds as long as the total vowel-recoverability term from the transition, jd_t, is greater for /sp/.

3.6 Task-specific effects and isochrony

In chapter 2, we noted that many previous experiments on compression effects have used reading tasks such as word lists or a single repeated carrier sentence. It was hypothesized that using a set of rhythmically identical stimuli may result in speech that is artificially isochronous and displays compression effects that are not characteristic of natural speech. Indeed, the less-isochronous speech elicited in the production experiment failed to display some of the compression effects

that have been found in earlier studies. For instance, Munhall et al. (1992) find incremental CS for obstruent clusters in coda position, a result that was not replicated in the current study.

We discussed a number of reasons that the results of these experiments may have come out different. One possibility is that there really was some incremental CS with obstruent-obstruent clusters in this study, but they weren't detected because the variance was higher than previous studies. If this or one of the other explanations discussed in chapter 2 is the correct one, than nothing further need be said about the differences. However, the isochrony explanation is a way to explain the differences between experiments while still accepting the validity of all reported results; as such, it is worth demonstrating that this explanation can in principle work. In this section, we illustrate how these differences might arise from effects that are specific to highly rhythmic, isochronous speech.

The facts to be accounted for are that obstruent clusters fail to display incremental CS in natural speech, but display incremental CS in more rhythmically-constrained speech. The analysis developed in the previous sections already accounts for the cases where no incremental CS is observed. So we require a theory of how additional CS effects could arise in rhythmically-constrained speech.

One account of these effects relies on a task-specific production constraint that acts upon the output of the grammar.[3] We can think of the outputs of the grammar, which are the input to a

[3] Note that we could equally well posit an isochrony constraint that is part of the grammar and affects the parameters (i.e., constraint weights) used by a speaker in any given situation. We

speech task, as a series of timing units that have relative durations assigned to them by the grammar, but do not have absolute durations, which depend on speech rate and sundry performance factors. Isochrony, then, can be conceived of as a manipulation of speech rate at various points in an utterance; this manipulation will create effects that look like compression, but don't have their roots in the grammar at all.

In this type of analysis, three repetitions of the carrier phrase 'please say X twice' with different target words might be represented as below:

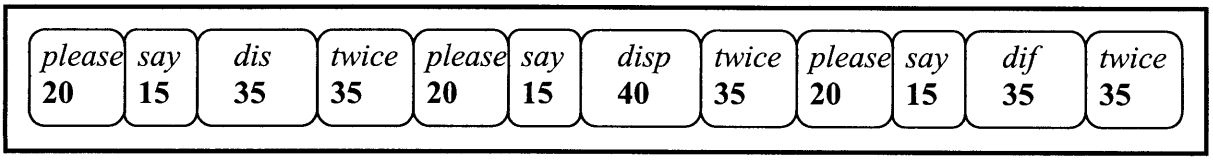

Figure 3.14. *Relative temporal representation of three repetitions of a carrier sentence with different target words. Boldface numbers represent relative duration in arbitrary units.*

The tendency toward isochrony in multiple repetitions can be characterized as a tendency to keep the temporal interval between each element in a sentence and its corresponding elements in the adjacent sentences equal. So, for instance, isochrony would favor keeping the interval between *please* in the first sentence and *please* in the second sentence equal to the interval between *please* in the second sentence and *please* in the third sentence. Exactly where the cognitively relevant interval begins and ends, and how it should be measured, is a controversial question that has no universally accepted answer. For concreteness, we'll assume that the relevant interval begins and

---

develop this extra-grammatical version instead because it is simpler and could in principle be applied to domains other than speech, such as text-setting or hand-clapping.

ends at the temporal midpoint of words; this is probably a simplification, but it won't affect the conclusions of the analysis. Virtually any characterization of the relevant interval will result in longer measurements when a longer target word is contained within that interval; this is the only necessary condition to derive the type of compression shown here.

When target words in various sentences are not assigned the same relative duration by the grammar, preserving isochrony will require some kind of adjustment to the speech stream. This is shown in figure 3.15, where unequal target words result in relative duration differences between consecutive intervals of the type relevant to isochrony.

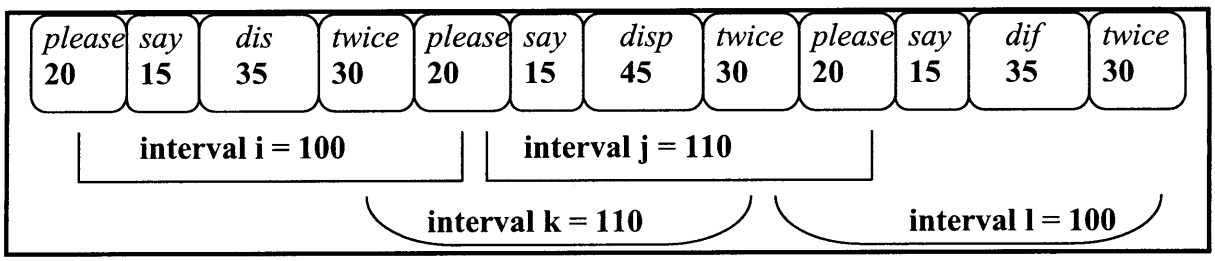

Figure 3.15. *Illustration of the intervals that isochrony acts upon. Isochrony would favor keeping interval i equal to j and k equal to l. Boldface numbers represent relative duration in arbitrary units.*

Let us assume that relative durations are turned into absolute durations when they are assigned a speech-rate coefficient, a real number that is multiplied by relative duration. This may be a simplification, but again, it won't affect our conclusions. Given this implementation of speech rate, the only way to preserve isochrony will be to either produce the longer target word *disp* at a faster speech rate or produce the shorter target words *dis* and *dif* with a slower speech rate. These

solutions are illustrated in figure 3.16. This figure assumes that the domain of speech rate manipulations is the word; larger or smaller units would work equally well.

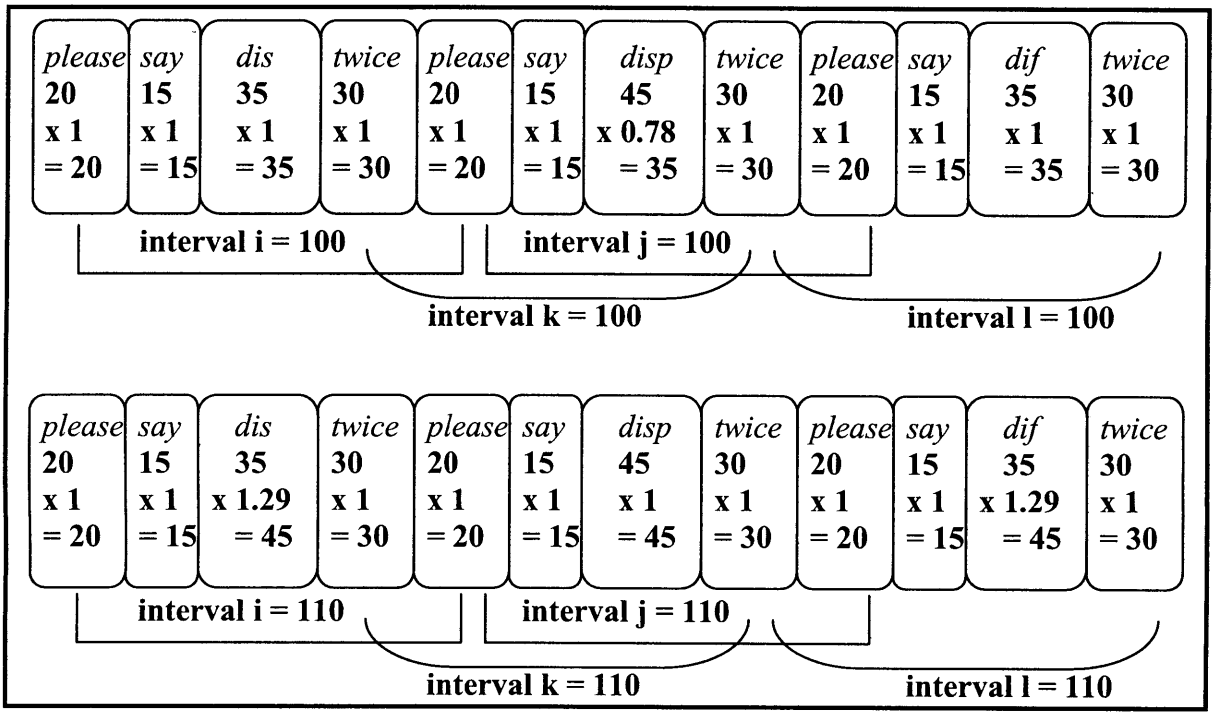

Figure 3.16. *Using speech rate to preserve isochrony by shortening the longer target word (top) or lengthening the shorter target words (bottom). Boldface numbers represent relative duration in arbitrary units, speech-rate coefficient, and absolute duration in arbitrary units, from top to bottom.*

The result of the isochrony constraint is that inherently longer target words are produced at a faster speech rate than inherently shorter target words. Simplifying again, we'll assume that the speech-rate coefficient applies equally to all parts of a word; the only crucial part of this assumption, however, is that speech rate has some effect on vowel duration. Given this theory of isochrony, we predict the following forms for *dis* and *disp* under isochrony:

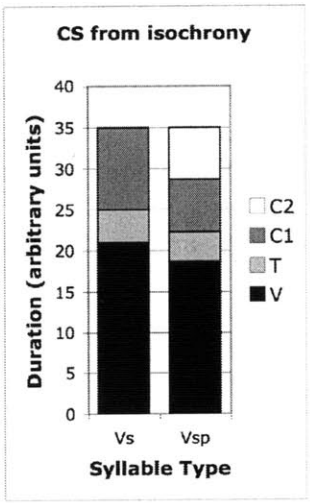

Figure 3.17. *Incremental CS induced by the isochrony constraint, in a case where the grammar alone would not predict it.*

The vowel, as well as all of the other components, in *disp* shorten to make it the same duration as *dis*. This is despite the fact that the grammar itself predicted no incremental CS for these two items.

This serves as a demonstration that overly rhythmic speech could induce compression effects that are not characteristic of more naturalistic speech. It suggests that we should be very cautious in drawing conclusions about temporal patterning from data that are elicited in such conditions. The implementation of speech rate and isochrony here is simplified, but the point still stands.

This approach derives dramatic, complete isochrony. In fact, the effects in the Munhall *et al.* study were small; most were on the order of 5-10 ms. We could model partial isochrony by expanding the formalism to include competing weighted constraints on isochrony. Such a model

would also be able to derive a tendency towards evenly-spaced words, but not complete isochrony. This might emerge from competition between the pressure to be isochronous and, for instance, a dispreference for sudden large changes in speech rate. Such a system would not fundamentally change the conclusions of this section, that the duration of a longer and shorter target word will be shifted so the two are closer to each other.

3.7 Conclusion

In this chapter, we've developed a formal, constraint-based analysis of compression effects in English. The analysis relies on the logic of the weighted-constraint system, whereby the phonetic form of linguistic items is shaped by trying to find the best compromise between a set of conflicting pressures on timing. For the case of CS, those pressures are duration targets for segments on the one hand and larger units on the other.

As more segments are 'crowded' into a syllable, the analysis predicts that all of the segments inside the syllable will shorten to some extent, and the syllable itself will lengthen to some extent. We also proposed a minimum inherent duration past which segments can not shorten under any circumstances. CS will not occur when segments are close to their minimum durations.

Asymmetries in CS between different types of segment, or the same segment in different contexts, are analyzed in two ways, both concerning perceptual properties. The hypothesis proposed here is that part of the duration requirements for a segment can be satisfied by portions

of the speech stream that are not contained in that segment itself. For instance, the duration target by a vowel can be satisfied by information about that vowel contained in an adjacent transition or consonant.

For the grammar to produce outputs that are qualitatively similar to the data from the production experiment, we needed to make a series of assumptions about the relative vowel transparency of several parts of the speech stream. Those assumptions are:

- Steady states of liquids contain more information about an adjacent vowel than steady states of obstruents.
- In onset position, nasal steady states or transitions or both contain more vowel information than those of obstruents, but not in coda position.
- Liquid steady states or transitions or both contain more vowel information in coda position than they do in onset position.
- Formant transitions following /sp/ clusters in onset position contain more vowel information than aspiration following /p/, or /s/ and the transient of /p/ contain more vowel information than just the transient of /p/, or both.

The analysis developed here relates these perceptual facts to compression effects: vowel steady states can shorten more when there is more information about them dispersed in the surrounding context. So, for instance vowels can shorten more next to liquids than they can next to obstruents.

If the vowel is already close to its minimum duration when adjacent to only one obstruent, we predict that there will be no incremental CS when a second obstruent is added. This is exactly what was reported in the production experiment from chapter 2. Because liquids contain more vowel information, the vowel will not be as close to its minimum duration when adjacent to a single liquid. When a second consonant is added, we do predict incremental CS. This is the fundamental asymmetry that was found in the production experiment.

Other asymmetries in compression, between nasals in onset and coda position, liquids in onset and coda position, and /p^h/ versus /sp/ items in onset position, were attributed to some combination of differences in the vowel-recoverability coefficients of the consonants 'proper' and the coefficients for transitions between consonant and vowel in the two positions. The analysis here is able to predict patterns of differential incremental CS based on either the consonants themselves or their transitions.

When coupled with a simplified model of isochrony and speech rate, the analysis here can also predict the differences between the current production study and previous studies. While the current study found no evidence for CS in obstruent clusters, previous studies have. Although there are several plausible explanations for why these differences might have arisen, we chose to explain them in terms that would allow us to accept the reported results of all studies, rather than positing erroneous effects stemming from methodological problems or missing effects due to variability.

The important difference is that those studies used more rhythmically repetitive tasks, such as word lists or a single short repeated carrier sentence. The analysis of the differences between studies relied on proposals about the extra isochrony involved in the earlier studies. When the outputs of the constraint system are fed into a separate, task-specific system for enforcing isochrony, compression effects may emerge that were not present in the output of the grammar itself. In other words, we can reduce the differences in results between experiments to a difference in tasks.

In the next chapter, we describe an experiment meant to test the perceptual hypotheses laid out here. The experiment requires subjects to identify a vowel based only on the adjacent consonant or on that consonant and part of the transitions to the vowel. If the asymmetries in compression discovered in chapter 2 turn out to be mirrored by asymmetries in perception, we should strongly prefer a theory that relates the two sets of facts, such as the one developed in this chapter.

Appendix 3A Equations for solving the three-segment optimization problem

We begin with the cost function:

(1) $$f(x, y, z) = w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma)^2 + w_2 \bullet (d_x - t_x)^2 + w_2 \bullet (d_y - t_y)^2 + w_2 \bullet (kd_y + jd_t + d_z - t_z)^2$$

The partial derivatives for x, y, and z, respectively, are as follows:

(2) $$f'_{y,z}(x) = 2w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma) + 2w_2 \bullet (d_x - t_x)$$

(3) $$f'_{x,z}(y) = 2w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma) + 2w_2 \bullet (d_y - t_y) + 2w_2k \bullet (kd_y + jd_t + d_z - t_z)$$

(4) $$f'_{x,y}(z) = 2w_1 \bullet (d_x + d_y + d_t + d_z - t_\sigma) + 2w_2 \bullet (kd_y + jd_t + d_z - t_z)$$

Solving out for z in terms of x and y:

(5) $$z = \frac{d_y \bullet (w_2 \bullet ((k+1) \bullet k + 1) + 2w_1) + 2w_1d_x + d_t \bullet (w_2j \bullet (k + 1) + 2w_1) - w_2t_y - w_2t_z \bullet (k + 1) - 2w_1t_\sigma}{w_2 \bullet (k+1) + 2w_1}$$

We use this solution to solve out for y in terms of x:

$$(6)\qquad y = \frac{\left[\begin{array}{l} w_2^2 \bullet (k+1) \bullet ((d_x - t_x) \bullet (k^2 + k + 1) - kt_y - jd_t + t_z) \\ + 2w_1w_2 \bullet (d_t \bullet (k \bullet (k+1) - d_t \bullet (k+2)) + d_x \bullet (2k^2 + 3k + 2) - t_x \bullet (k^2 + 2k + 2) - kt_y + t_z \bullet (k+2) - kt_\sigma \bullet (k+1)) \\ + 4w_1^2 \bullet (d_x \bullet (k+1) - t_x + d_t \bullet (k-j) - t_x + t_z - kt_\sigma) \end{array}\right]}{2w_1w_2 \bullet (k^2 - 2) - w_2^2 \bullet (k+1) - 4w_1^2 \bullet (k-1)}$$

Finally, we use the expressions in (5) and (6) to solve for x solely in terms of the constants. The solution is:

$$(7)\qquad x = \frac{(2w_1w_2^2 \bullet (k+1) + 4w_1^2w_2) \bullet (d_t \bullet (1-j) + t_y \bullet (1-k) + t_z - t_\sigma) + t_x \bullet (2w_1w_2^2 \bullet (k^2 - k^3 - 4) - 2w_2^3 \bullet (k+1) + 4w_1^2w_2 \bullet (2k - k^2 - 2))}{2w_1w_2 \bullet (k^2 - 2) - w_2^2 \bullet (k+1) - 4w_1^2 \bullet (k-1)}$$

4 An experimental investigation of vowel recoverability from consonants

4.1 Introduction

4.1.1 Preliminaries

The experiment in chapter 2 found that patterns of incremental CS differ across consonants. Liquids drive incremental CS, while obstruents do not. Nasals drive incremental CS in onset position but not in coda. For liquids, the incremental CS effect is slightly larger in coda than in onset position. In chapter 3, we developed a grammar that is capable of characterizing these asymmetries; that grammar relied on several assumptions about the perceptual properties of consonants and transitions. In this chapter, we describe an experiment that will test those perceptual hypotheses. The general finding is that patterns of compression in language production mirror asymmetries in speech perception.

The hypothesis put forward to explain compression asymmetries is that the amount of vowel shortening allowed in any context depends on how much perceptual information about that vowel is present in the context itself. For instance, we hypothesized in the preceding chapters that liquids contain more information about an adjacent vowel than obstruents do; for this reason, the interval of 'pure' vowel that is not overlapped with the adjacent consonant can shorten more next to a liquid than next to an obstruent.

The experimental hypotheses to be tested here, then, have to do with differences in relative sensitivity to vowel contrasts between cases where the surrounding context is an obstruent and cases where it is not. Although a fair number of researchers have studied vowel identification

from adjacent obstruent noise, we know of no previous studies that have investigated subjects' ability to identify vowels from an adjacent nasal or liquid.

The general hypothesis to be tested is that patterns of CS can be explained by the distribution of vowel information over time. When adjacent segments contain more information about a vowel, the steady-state of that vowel can shorten more; the adjacent context helps satisfy the duration target of the vowel. Patterns of CS in production should be mirrored by patterns of sensitivity in perception.

Based on the asymmetries in production discovered in chapter 2, we constructed a grammar in chapter 3 that generated the following predictions about perception:

- Steady states of liquids contain more information about an adjacent vowel than steady states of obstruents.
- In onset position, nasal steady states or transitions or both contain more vowel information than those of obstruents, but not in coda position.
- Liquid steady states or transitions or both contain more vowel information in coda position than they do in onset position.
- Formant transitions following /sp/ clusters in onset position contain more vowel information than aspiration following /p/, or /s/ and the transient of /p/ contain more vowel information than just the transient of /p/, or both.

In this chapter we attempt to test these predictions. The experimental paradigm used here is identification of forward- and backward-gated stimuli. Utterances of the same vowel are

recorded adjacent to several consonants of interest. The consonants are then extracted from recordings and played to subjects without the adjacent vowel. Subjects are asked to identify which word these truncated stimuli came from, which involves an implicit identification of the vowel. In addition, successive 'gates' add back in small intervals of the transition between vowel and consonant, making the task easier at each successive gate. Examining the incremental increases in sensitivity at each gate allows us to test hypotheses about the amount of vowel information in transitions.

One difficulty that arises in interpreting the results of the experiment pertains to how gross, global hypotheses about the 'vowel transparency' of various items ought to be reflected in binary-choice identification data from specific pairs of vowels. Each vowel in a language, of course, contrasts with a number of other vowels; it is not necessarily the case that all of these contrasts are affected in the same way by differences in the quality of an adjacent consonant. When we say that liquids contain more vowel information than obstruents, what exactly does it mean in perceptual terms?

Given that consonantal differences may have different effects on different vowel contrasts, it seems unlikely that statements about relative vowel information should hold for every single vowel contrast in the language. Even if we could test every contrast, which would be an enormous task, it's plausible that we would find different effects for different contrasts.

In the absence of a perfect characterization of the function from gross vowel perceptibility to contrast-specific sensitivity, we will work with the assumption that something like 'a

preponderance of the evidence' from various vowel pairs should agree with our predictions before we count them as confirmed. In practice, this is a fuzzy and relative notion: the more vowel pairs that a generalization is valid across, the more confident we can be in that generalization.

We predict, then, that subjects should show more sensitivity to liquid stimuli than to obstruent stimuli in both onset and coda position, in the condition where only the consonant is played to them (referred to as the *zero gate*). This will be easy to test by simply examining the data from the zero-gate condition. We expect sensitivity to vowel contrasts to be higher with liquid stimuli than with obstruent stimuli.

The theoretical model in chapter 3 also made a number of predictions that may hold of the consonant steady state, the transition, or some combination of the two. For instance, it was hypothesized that the steady-state of a vowel following a /sp/ cluster may be shorter than that following /p^h/ because the formant transitions in the /sp/ case are more 'valuable' than aspiration, in the sense of contributing more to vowel perceptibility. Alternatively, the duration asymmetry may hold simply because the /s/ contains information about the following (non-adjacent) vowel. And of course, the duration effect could follow from some combination of these two putative perceptual effects.

The current study also examines two gates where a portion of the transition between vowel and consonant is included in the stimuli. This allows us to test for any large differences between segments in the increment to sensitivity given by the transition. We expect, then, that for each

vowel pair, either sensitivity at the zero gate should be higher for /sp/ than /p/ items, or the transition increment in sensitivity associated with /sp/ items should be larger than for /p/, or both.

Similarly, we hypothesized differences between onset and coda position in the 'value' of either transitions or steady states for /n/ and liquids. We predict that zero-gate sensitivity or transition increases or both should be greater in coda than in onset position for liquids. For /n/, which patterned with obstruents in coda position (no incremental shortening) but did show shortening in onset position, we predict that zero-gate sensitivity or transition increases or both should be greater than those for /p/ in onset position, but not coda position.

4.1.2 Previous studies

Several previous experiments have shown that subjects are able to identify vowels at a level above chance from adjacent obstruents alone. These studies have used both gating and 'silent-center' stimuli, where some or all of the vowel in a CVC word is removed. Here I summarize the findings and note a few analytical issues that figure prominently in our analysis of the results.

In English, subjects identify vowels at a level above chance from both preceding and following voiceless stops (Winitz *et al.* 1972). The preceding stops included aspiration; the following ones consisted only of the burst. They also perform above chance with whispered transients, not including frication, from preceding voiced stops (Repp & Lin 1989). Subjects identify vowels at a level above chance from preceding (Yeni-Komshian & Soli 1981) and following (Whalen 1983) sibilant fricatives, both voiced and voiceless. Whalen reports that subjects are above

chance at discriminating rounding contrasts and height contrasts. Nine of the ten subjects have higher percent correct for roundness than for height.

Silent-center studies, where almost the entire vowel is excised from CVC stimuli, also provide relevant data. Parker & Diehl (1984) report that subjects perform above chance with /dVd/ stimuli that have 90% of the syllable duration removed, and replaced with either silence or broadband noise. Rogers & Lopez (2008) report above-chance identification with /bVb/ stimuli that only preserve 10 ms after the initial burst and before the final closure.

The same type of results are reported for a few languages other than English. Krull (1990) reports above-chance vowel identification from preceding voiced stops in Swedish. Bonneau (2000) reports above-chance vowel identification from preceding voiceless unaspirated stops in French. Smits *et al.* (2003) and Warner *et al.* (2005) report that subjects show good discrimination of height and backness contrasts from the first third of a vowel, above 60% TI (a sensitivity measure that ranges from 0 at chance to 100 at perfect disrimination) in Dutch. For CV sequences, subjects appear to identify the vowel at a level above chance by the time they hear 2/3 of the preceding consonant, if not sooner.

Some of these studies, though not all, appear to show a *ceiling* effect when parts of the excised vowel are added back into stimuli. At some point, subjects reach maximum sensitivity (close to 100% correct), and adding more vowel material back into the stimuli generates diminishing returns. For studies that report relevant data, it appears that the ceiling tends to occur within the first 40% of the vowel's duration.

4.1.3 Reanalysis of previous studies

The studies discussed above have shown that subjects can identify vowels based on adjacent obstruent noise alone. Given that the current experiment will attempt to extend these findings to other consonants, and will require choices about which vowels to examine, it would be useful to know how sensitive subjects are to various vowel contrasts. With the exception of the Warner *et al.* (2005) study on Dutch, however, the analyses in these papers are not set up in a way that allows us to conclude anything about sensitivity to contrasts. We digress to discuss the analytical issues in greater detail, because they apply to the current study as well.

The problems stem from two related conceptual issues: *bias* and *sensitivity*. Roughly speaking, these studies fail to distinguish between the likelihood of responding to some stimulus α with response α and the likelihood of responding α *in general*; this is the issue of response bias. In addition, these studies fail to distinguish between subjects' *accuracy for a given category* and *sensitivity to a given contrast.*

All of the statistical analyses in these papers, with the exception of Whalen's, ignore the question of bias completely when they analyze data. If they find that subjects respond α relatively often to stimulus α, they conclude that α is easy to identify. In reality, we don't know how much of these effects are attributable to properties of α stimuli until we compare how often subjects respond α to *non*-α stimuli. Factoring out bias is a crucial preliminary to learning about similarities and differences between stimuli.

Even after factoring out bias, it doesn't make a lot of sense to talk about accuracy for a given category. Surely, a subject's likelihood of correctly responding α to an α stimulus depends in part on what the other possible responses are. For instance, in experiments that use vowel sets such as {i, a, u}, we generally find that accuracy is very high for /i/ stimuli. But when we add in vowels such as {e, ɛ, ɪ}, this effect disappears. What these results are telling us is not that /i/ is more identifiable as a category than other vowels; they are telling us that [i-a] and [i-u] are more distinct contrasts than [u-a], or that there is a bias to respond /i/ more often than /u/ and /a/, or some combination of the two. All identification errors are not equal and are not generally equally likely; the likelihood of correctly identifying a stimulus depends in part on a subject's sensitivity to the contrasts that involve that stimulus. It doesn't make sense to attribute sensitivity to a category; sensitivity is a property of contrasts.

To learn about sensitivity to contrasts, we must construct a model that distinguishes sensitivity from bias. Toward that end, some results from three of the studies reviewed here were reanalyzed: Whalen 1983, Parker & Diehl 1984, and Repp & Lin 1989. These studies either provided raw count data or provided enough detail that count data could be reconstructed. For the first two studies, those data were analyzed using a hierarchical log-linear regression model. The model attempts to predict the log frequencies of each stimulus-response pair by fitting parameters that represent relative bias for each category present in the experiment and sensitivity to each contrast present in the experiment. Because it wasn't possible to reconstruct data for each individual in the experiments, these models inflate the number of observations and consequently the probability of Type I error (rejecting a true null hypothesis); however, they at least provide us with an account of the data that takes bias and sensitivity into account. The Repp & Lin study

was reanalyzed using Luce's (1963) Biased Choice Model, which also distinguishes between bias and sensitivity. Significance tesing was not carried out for this data set; rather, the distance and bias parameters of the model were examined to confirm that they are consistent with the other experiments. Appendix 3A contains a detailed description of each of the reanalyzed experiments.

The general finding that subjects are able to tell apart some vowels based only on their surrounding contexts at a level above chance still stands; in fact, this finding shouldn't be affected by bias or sensitivity anyway. The only possible exceptions are 'one-step' height contrasts, contrasts between vowels that differ only in being high as opposed to mid. In the Whalen study, which used only fricative noise, subjects are not significantly above chance for the [u-o] contrast. In the Parker & Diehl study, which included a few periods from each edge of the vowel, subjects do appear to be significantly above chance for [ɪ-ɛ], but are significantly more sensitive to the [ɪ-ʌ] and [ɛ-ʌ] contrasts. In the Repp & Lin study, distances parameters for one-step height contrasts are by far the lowest; four out of 18 are actually slightly negative, indicating below-chance discrimination.

Subjects are more sensitive to backness/rounding contrasts than they are to height contrasts. In the Whalen study, [i-o] and [i-u] are the two most discriminable contrasts. When the effects of rounding, height, and the combination of the two features are taken into account, the independent effect of rounding is significant but the independent effect of height is not. In the Repp & Lin study, backness/rounding contrasts, with the exception of [æ-ɑ], are among the most

discriminable in the experiment. Two-step height contrasts such as [i-ɛ] and [e-æ] are nearly as discriminable in the context of /d/ and /b/, but much less discriminable in the context of /g/. Acoustic analysis provided by the authors suggests that this is probably due to the fact that /g/ transients show extensive, even allophonic, coarticulation along the backness/rounding dimension, while the height dimension is compressed.

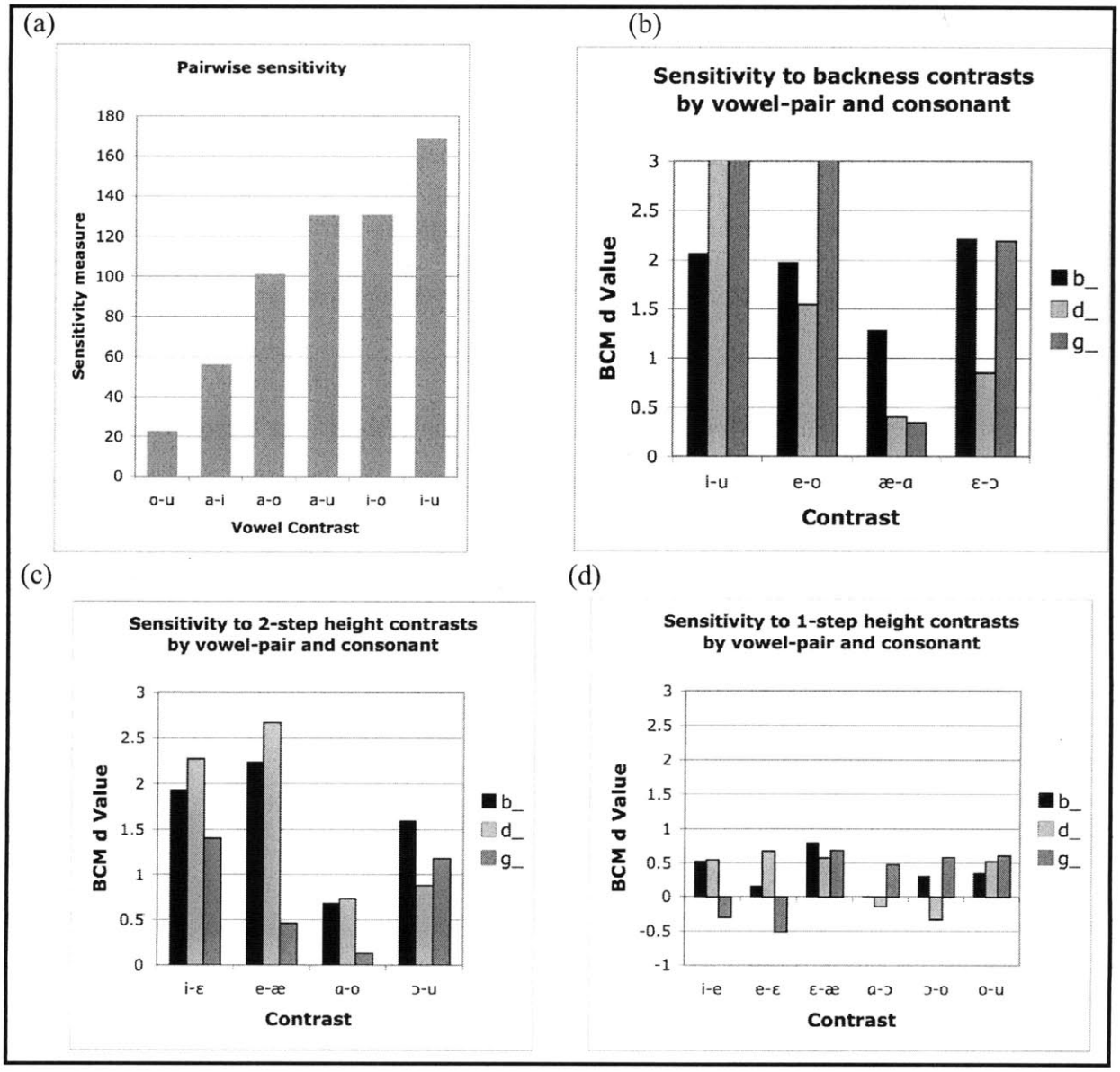

Figure 4.1(a). *Distance/sensitivity parameters fit to Whalen's (1983) data by a log-linear model.* **(b-d).** *Distance parameters for a subset of Repp & Lin's (1989) data, derived from the Biased Choice Model. Graphs show d values for vowels differing in one step along the height dimension (b); differing along the front/back dimension (c); differing in two steps along the height dimension (d). The data in (b-d) are organized by preceding consonantal context. Bars that go off the top of the chart represent stimuli that are never confused; they have arbitrarily high d values.*

Figure 4.1 shows some characteristic sensitvity data from two of the studies. Subjects tend to be more sensitive to contrasts involving rounding than those not involving rounding; this may suggest that obstruent noise carries more cues to rounding than to other contrasts. One-step height contrasts, which generally don't involve rounding, are the most difficult to discriminate.

These studies provide some useful lessons for constraining the design of the current experiment. They show that subjects can tell the difference between many vowels using only the information present in an adjacent obstruent. They are most sensitive to contrasts in backness/rounding and height contrasts that involve more than one step along this dimension; they are less sensitive to contrasts that involve only one step along the height dimension. When portions of the vowel are added back into the signal, they reach maximum sensitivity (close to 100% correct) sometime in the first 40% of the vowel.

4.2 Methods

We constructed pairs of stimuli that differed only in their vowels: the vowel pairs tested are [e-o], [i-e], and [ɑ-u]. The idea was to check a small number of vowel contrasts that represent the different types examined in prior studies: one differing along the backness/rounding dimension, which is generally found to be the most discriminable type of contrast; one differing in more than one step along the height dimension as well as rounding, which should be roughly comparable to the backness/rounding contrast; and one differing only in one step on the height dimension, which is generally found to be the least discriminable type of contrast.

Stimuli consisted of all combinations of the relevant vowels with the consonants {r, l, n, p, sp} in onset position and {r, l, n, s} in coda position, matching the consonants tested in the production study. A few stimuli were excluded: the pair [es-os], because there is no single onset consonant that could combine with both sequences to make a word; and the sequence /ur/, due to its dubious phonotactic status. To replace [es-os], we included [ep-op]; although it is probably not the case that /p/ and /s/ contain the same vowel information, the /p/ will at least be comparable to onset stimuli.

Two native speakers of American English from eastern Massachusetts (1 female, 1 male) were recorded producing three repetitions of each stimulus item in the carrier sentence 'I bet ___ is the answer'. All recorded materials were segmented following the procedures of the production experiment, detailed in sections 2.1-2.2 and appendix 2A. One token of each stimulus from each speaker was selected for inclusion in the experiment. For each item, the selected token was the one with consonant and VC/CV-transition durations closest to each subject's mean for the item.

The selected tokens were segmented into several gated stimuli. The first one, referred to as gate 0, contains only the acoustic steady state of the consonant, with none of the transition to or from the vowel. Succeeding gates incrementally added 20-27 ms. of the VC/CV transition and, in some cases, vowel steady state (the shortest transitions in the experiment were 35-40 ms). For any given vowel pair, the gate durations were chosen so as to be maximally close to the marked boundary between the transition and the vowel steady-state for those tokens where this consideration was relevant. For instance, in the [i-e] onset condition, the shortest transitions

clustered in the 45-50 ms range (for obstruents and /n/); a gate duration of 23 ms was used, meaning that the second-gate stimulus is truncated within 5 ms of the marked transition-vowel boundary.

The end result is that all stimuli (across consonants) within each crossing of vowel-pair and syllabic position have the same gate duration, but the gate duration varies slightly between vowel pairs. The stimuli were truncated at the zero-crossing closest to the chosen gate duration; this resulted in differences of up to 2 ms in gate duration between items in the same condition. Some of the stimuli that included stops were edited to remove a noticeable electrical buzz from the closure portion of the recording. The figure below shows a pair of recordings used to derive stimuli for the experiment, and the segmentation strategy for creating those stimuli.

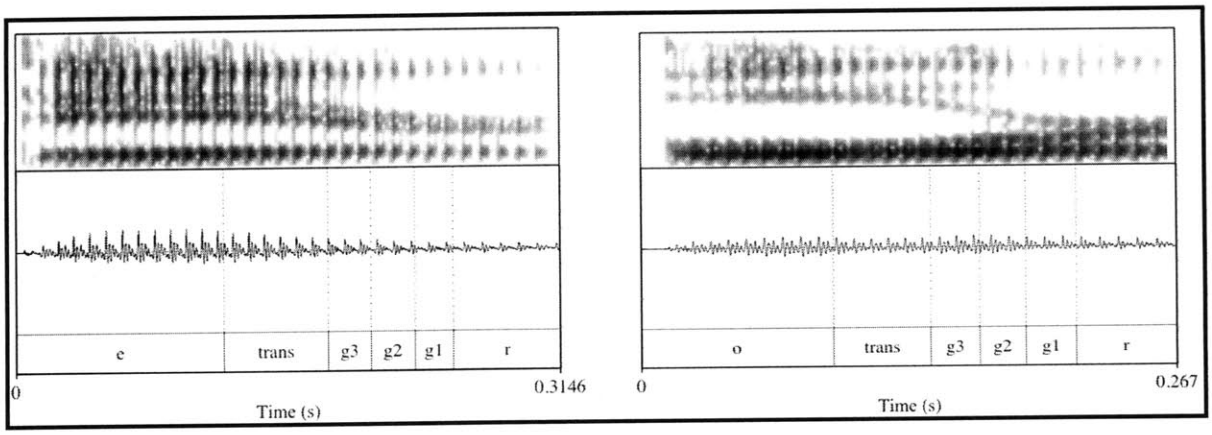

Figure 4.2. *Tokens of /er/ (left) and /or/ (right) used in the experiment. Text grid shows three gates taken from the right edge of the vowel-consonant transition. Non-gate segmentation based on F3. 0-gate stimuli would consist of only the portion marked 'r' here; successive gates would add the segments labeled 'g' above to that original 'r', one 'g' section at each gate. Those gate*

sections are part of the transition; in this case, even the longest gate (number 3) wouldn't include the entire transition. The 'left over' part of the transition is labelled 'trans'.

The intensity and F0 of the stimuli were not equalized in any way. Any differences between segments in these respects may themselves affect the process of determining the quality of an adjacent vowel, and eliminating differences could alter the identification results in ways that don't reflect natural speech.

Impressionistically, the sounds were rather easy to distinguish by the second gate, 40-55 ms into the transition. Short pilot studies were conducted for each vowel pair using gates 0, 1, and 2. The results indicated that most subjects obtained 80-90% accuracy by the second gate. At gate 0, accuracy ranged from slightly below chance to around 70%, depending on subject and stimulus. Subjects performed around chance at all gates for the reverse-gating (coda consonant) [i-e] condition; this is presumably because /e/ is followed by an offglide that is nearly identical to /i/. This condition was dropped from the study.

Due to the large number of stimuli, and the difficulty of focusing on an identification task for long periods of time, the stimuli were split into five groupings that we refer to as separate conditions. Each subject participated in one of these conditions. Each condition examined a single vowel pair with either onsets or codas. Each block consisted of one stimulus from each speaker, with each consonant-vowel pair, at each gate. In the onset conditions, for instance, each

block would cross two speakers, two vowels, six consonants,[1] and three gates, for a total of 72 trials. As each audio file was played, a choice of two words appeared on the screen; subjects used a left and right button to identify the corresponding word as the one they had heard part of. Subjects were given 1 second to respond; after this, the message *Timeout!* appeared at the center of the screen for 300 ms. Stimuli were randomized within each block; subjects were given the option of taking a break after each block except the first. The first block consisted of training without feedback, with gate 3 of each stimulus item (containing more transition/vowel content than any of the actual test items) played once. Impressionistically, the design was rather fast-paced and tended to be surprising at the beginning; the training block was included for this reason.

All word choices were existing lexical items of English; this sometimes required an orthographic consonant that wasn't present at all in the auditory stimulus. For instance, subjects were played a fragment of /ep/ and asked whether it was *cape* or *cope*, despite the fact that there was no hint of a /k/ in the recording. Wherever possible, the choice of this 'fixed consonant' was held constant across target consonants within each vowel pair (e.g. *care-core, kale-coal, cane-cone, cape-cope*); in a few cases this wasn't possible. Word pairs were not balanced for frequency; this would probably have been impossible given the nature of the task, and we can correct for frequency effects by interpreting the results with a statistical model that separates the effects of bias from the effects of similarity. Lexical bias, for instance, might lead subjects to respond with

[1] One consonant in addition to the five mentioned above, /s/, was present in the onset experiment but was not analyzed in the end, because it was not directly relevant to the production patterns from the previous experiment.

knee more often than *neigh*, but this would show up in the statistical model only as increased bias toward /i/ in the context of /n/, not as increased sensitivity to the contrast.

For the [e-o] onset condition, 15 subjects were tested. For the [ɑ-u] coda condition, 10 were tested. For the other three conditions, 11 subjects were tested. All reported being native speakers of American English who had never been diagnosed with any speech, hearing, or reading disorders. All subjects were compensated for their time. The tests were run in the Behavioral Research Lab at MIT, with up to 10 subjects simultaneously at workstations separated by dividers.

The results were analyzed using a logit mixed effects model, implemented with the lme4 package (Bates 2007) in the statistical platform R. This model is similar to the linear mixed effects model discussed in chapter 2, except that it attempts to model binary, categorical data in terms of the binomial distribution. The model is fit using the Laplace approximation. The dependent variable was one of the two vowel responses. Random effects were speaker and listener. The model included a fixed effect for each stimulus vowel, each consonant, and the interactions between them. In such a model, the effects that correspond to sensitivity will be those that include a term for a stimulus vowel. For instance, the effect of 'stimulus /o/' in the [e-o] condition, where the dependent variable is 'response /o/', will tell us how much more likely subjects are (in log odds) to respond with /o/ when the stimulus is /o/ than when the stimulus is /e/. Further fixed effects included whether or not each trial followed an error on a previous trial, whether it followed a timeout on a previous trial, and the number of trials that had passed since the beginning of the experiment. Adding trial number to the models resulted in singular

convergence for the model-fitting algorithm, indicating that the data is not complex enough to justify a model with a separate variable for trial number. This variable was therefore excluded from subsequent models.

Separate models were constructed for the zero-gate data and the transition data. After constructing a baseline model as described above, variables corresponding to sensitivity terms were removed from the model if they were clearly not significant, with a p-value greater than 0.1. This allows us to generalize about sensitivity to different categories of contrast; it also makes the model easier to fit. Variables were added to each model to check whether within-manner differences between consonants (e.g. /l/ vs. /r/) were significant. By-subject slopes were then added to the model, to capture differences in bias and sensitivity between subjects and between speakers. As there were more subjects than speakers (just two in the latter case), and variation between subjects was much greater than that between speakers, by-subject effects were tested first.

The second model, which examined the increase in sensitivity when transitions were added back into the zero-gate stimuli, was somewhat more complicated. This models tested specific hypotheses about differences in 'transition increments' to sensitivity across combinations of consonantal manner and syllable position. The modeling routine was identical to that described above, except fixed-effect interactions were added for the second gate and the particular lexical item presented was included as a random effect. This allows us to test whether the increase in sensitivity between gates in one condition is significantly different from another condition. This model compares differences between differences at many levels of recursion. For instance, we

might start with the difference between sensitivity to stimuli with /n/ and stimuli with singleton obstruents; then ask if that difference is larger at the second gate then at the zero gate; then ask if that difference between differences is larger in onset than coda position. Recall as well that sensitivity itself is equated with a difference in response likelihoods across two conditions. As such, the effects of interest are often interactions of relatively high order. This is a logical consequence of the fact that we're interested in differences in the way that segments and context interact with contrasts, rather than subjects' absolute ability to tell one vowel from another in adverse conditions.

Significance-testing is complicated in logit mixed-effects models. The lme4 software package returns a Wald *Z* statistic, which can be used to derive a *p*-value. However, there is some concern that this method is anti-conservative, tending to increase the probability of Type I Error. An alternative approach, if one is comparing hierarchically nested models, is to perform a chi-square test of the likelihood ratio between models with and without the relevant level of the variable; this approach is taken by Bates (2008), for example. This method generates higher p-values than those associated with the *Z* statistic, suggesting it is less anti-conservative than that method. Statistics reported here come from the likelihood ratio test. Fixed effects will be reported with an effect size β, representing the change in log odds associated with that effect; a chi-square statistic from the likelihood-ratio test, and a p-value from that test. Random effects, which are also evaluated with a chi-sqare test of likelihood ratios, are reported with just the latter two values.

4.3 Results

4.3.1 Zero-gate stimuli

Among the zero-gate stimuli, those that include no vowel or transition, there is generally higher sensitivity to vowel contrasts for stimuli containing liquids than any other kind. Stimuli with /n/ and /p/ induce the lowest sensitivity to vowel contrasts, and stimuli with /s/ and /sp/ induce an intermediate level of sensitivity.

The figure below shows sensitivity to vowel contrasts across stimuli with different manners of consonant; the sensitivity parameters were fit by a logit mixed model. The data are averaged across all factors except for consonant manner; as such, they fail to show some large differences between conditions. Those differences will be discussed below, but we briefly consider gross patterns of sensitivity first.

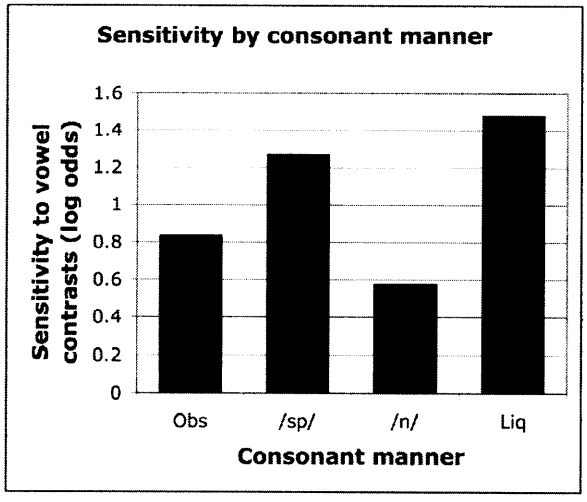

Figure 4.3. *Sensitivity to zero-gate stimuli by consonant, averaged across subject, speaker, and condition. The vertical axis shows the sensitivity parameter fit to each contrast by the model, in*

terms of differences in the log odds of a given response across stimulus categories. 'Obs' = singleton obstruent (/s/ or /p/ depending on condition); 'Liq' = liquid.

The descriptive results here bear on two of our experimental hypotheses. They broadly confirm the hypothesis that liquid steady states contain more vowel information than other segment types. And they suggest that /sp/ clusters contain more vowel information than singleton obstruents (although a tightly-controlled comparison to aspirated stops in onset position won't come until later in this section). Recall that we also posited a possible asymmetry betwenn /sp/ and /p/ pertaining to transitions; we test this below.

To test the other experimental hypotheses, we will need to examine vowel sensitivity as a function of the consonant in the stimuli and the particular condition. Shown below are the results across conditions. Note that some stimuli are not distinguished in this graph, because the model did not include parameters to distinguish between them. These were cases where collapsing hierarchically (e.g. one parameter for coda position rather than separate ones for [e-o] in coda and [ɑ-u] in coda) did not significantly decrease the fit of the model, i.e., cases where it was appropriate to generalize across related experimental items.

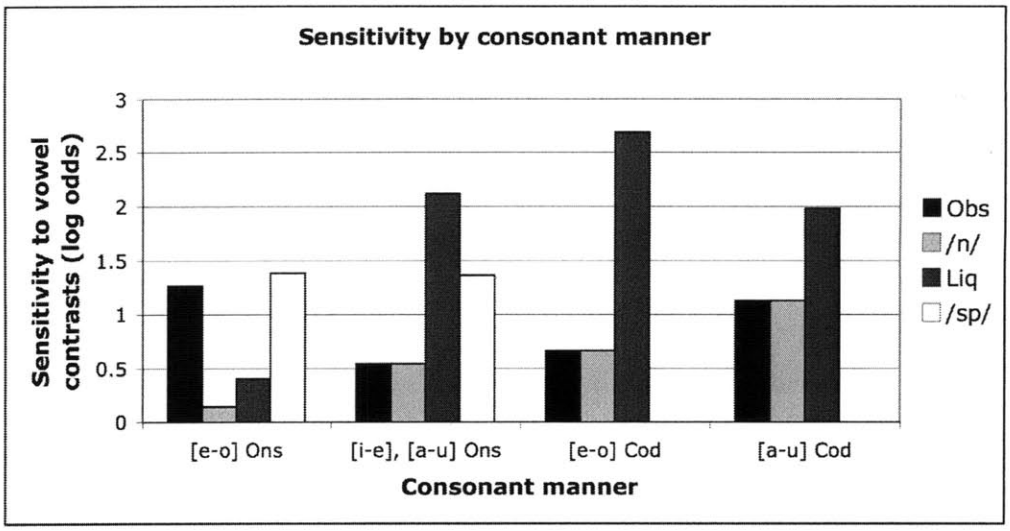

Figure 4.4. *Sensitivity to zero-gate stimuli by consonant and condition. The vertical axis shows the sensitivity parameter fit to each contrast by the model, in terms of differences in the log odds of a given response across stimulus categories. The name of each condition consists of the two vowels tested in the condition followed by 'Ons' for consonants in onset position or 'Cod' for coda. Only parameters that significantly improved the model fit are reflected here; contrasts corresponding to the other parameters are collapsed. The only exception is the difference between liquid stimuli in [i-e] and [ɑ-u] onset conditions, which is significant but is averaged here for visual ease.*

Figure 4.4 shows that patterns of relative sensitivity are broadly similar across all conditions except for [e-o] onset, represented by the leftmost series of bars. Except for that condition, subjects are most sensitive to vowel contrasts in stimuli with liquids. Sensitivity to stimuli with singleton obstruents and /n/ is statistically indistinguishable. Sensitivity to stimuli with onset /sp/ is higher than stimuli with onset /p^h/.

All of these patterns are different in the [e-o] onset condition. Here, sensitivity to items with obstruents (singleton and cluster) is highest. Sensitivity to stimuli with /n/ and liquids is extremely low; performance with /n/, in particular, is not significantly above chance.

As noted above, distinctions between sensitivity parameters were removed from the model in hierarchical fashion if they did not significantly contribute to the fit. This was done because it makes it easier to see generalizations that hold across multiple related items, and because it allows the model-fitting algorithm to run faster and converge in fewer iterations. The latter point is relevant because there are a large number of distinctions to be tested in this study and a large number of data points for each item; fitting the models discussed here, especially those with more random effects, is time-intensive.

The final model makes no distinction between sensitivity parameters in the [i-e] and [ɑ-u] onset conditions, except for liquid stimuli. It makes no distinction between sensitivity parameters for singleton obstruents and nasals, except in the [e-o] onset condition. Finally, it makes no distinction between sensitivity parameters for items containing /l/ and /r/; they display the same pattern with regard to other consonant manners, and adding a separate level of manner-dependent sensitivity to distinguish them does not significantly improve the model fit.

Subjects are more sensitive to vowel distinctions from stimuli involving liquids than stimuli involving obstruents in four of five conditions. In onset position before [ɑ-u], the difference is significant: $\beta = 2.18$, $\chi^2 = 45.1$, $p < 0.01$; we refer to this as the *baseline effect.* In onset position before [i-e], the difference is significantly smaller than the baseline effect: $\beta = -1.21$, $\chi^2 = 10.4$, p

< 0.01. In coda position following [e-o], the difference is significantly larger than the baseline effect; $\beta = 1.16$, $\chi^2 = 6.7$, $p < 0.01$. In coda position following [ɑ-u], the difference is significantly smaller than the baseline effect: $\beta = -1.32$, $\chi^2 = 11.2$, $p < 0.01$. In onset position before [e-o], the effect is reversed: subjects are more sensitive to vowel contrasts from stimuli involving obstruents. This reversal of the baseline effect across conditions results in a significant three-way interaction: $\beta = -3.04$, $\chi^2 = 43.8$, $p < 0.01$. In the two conditions where the difference is significantly smaller than the baseline, there is still a large effect in the expected direction.

Subjects are significantly more sensitive to vowel contrasts from stimuli with onset /sp/ than stimuli with onset /p/ in at least two of the three relevant conditions; the third is ambiguous. In onset position before [i-e] and [ɑ-u], the difference is significant: $\beta = 0.815$, $\chi^2 = 11.3$, $p < 0.01$. In onset position before [e-o], the effect is somewhat smaller; this interaction does not reach statistical significance in the final model.[2] This means that two of the three conditions show significantly better performance on stimuli with /sp/ than stimuli with /p/, and there is no clear evidence that the third condition differs from them, although the advantage for /sp/ is somewhat smaller in that condition.

The difference in sensitivity between items with singleton obstruents and items with /n/ is not significant in any of the conditions except [e-o] onset. Here, subjects are significantly less

[2] This interaction was near-significant before by-subjects effects were added into the model (p-value of 0.03 with Bonferroni-corrected α of 0.0125); it was retained for this reason. After accounting for subject variability with regard to the /p/-/sp/ comparison, however, the effect of this interaction shifted to become clearly non-significant ($p > 0.05$ with $\alpha = 0.0125$).

sensitive to vowel contrasts from stimuli with /n/ than stimuli with /p/: $\beta = -1.12$, $\chi^2 = 21.8$, $p < 0.01$.

Subjects differ on their overall accuracy, as well as their relative accuracy for liquid and /sp/ items, respectively, as compared to the other items in the experiment. Adding these differences into the model as by-subject random slopes significantly improved the fit, as measured by a chi-square test of likelihood ratios. For overall accuracy: $\chi^2 = 80.7$, $p < 0.01$. For sensitivity to vowel contrasts from items with liquids: $\chi^2 = 24.9$, $p < 0.01$. For sensitivity to vowel contrasts from items with /sp/: $\chi^2 = 20.4$, $p < 0.01$.

Some subjects essentially couldn't perform the zero-gate task. The subject with the largest negative random slope, for instance, identified 49% of the zero-gate stimuli correctly; chance performance is 50%. The subject with the largest positive intercept, in contrast, correctly identified 69% of the zero-gate stimuli. Most subjects lay in between these two extremes.

Subjects also varied in how much of an advantage stimuli with liquids had over stimuli with singleton obstruents. If we take the grand average for this parameter from the first model used above, which ignored differences in sensitivity for condition, as a rough guide, it suggests that 55 of the 58 subjects showed an advantage for items with liquids.

Similarly, subjects differed with respect to the advantage of /sp/ over /p/ items. We take the fixed effect sizes for the various conditions as a baseline to examine whether individual subjects showed the effect or not. For the [i-e] and [a-u] onset conditions, where the effect was largest, 21

out of 22 subjects showed an advantage for /sp/. For the [e-o] onset condition, only 8 out of 15 subjects showed the effect; this is why adding in subject effects changed the value of the fixed effect comparing these conditions. The remainder of the subjects in the [e-o] condition had either no effect or the opposite one.

Overall sensitivity to stimuli produced by the female speaker was somewhat greater than for the male speaker. Adding this difference into the model significantly improves the model fit: $\chi^2 = 25.2$, $p < 0.01$. The difference, averaged across all stimuli, is on the order of 0.4 logits. This could mean that the two speakers produced systematically different stimuli, or it could be an idiosyncratic property of the particular tokens that were recorded.

Finally, there was one significant task-related effect. Subjects performed significantly worse on trials following an incorrect answer on the previous trial; in other words, errors tended to come in bunches: $\beta = -0.23$, $\chi^2 = 10.1$, $p < 0.01$. This suggests that subjects may sometimes be aware when they answer incorrectly and that this may throw off their next trial. There was an effect of similar magnitude and in the same direction for trials following a timeout, a failure to answer on the preceding trial. This effect had much higher standard error associated with it, however, and did not reach statistical significance. This may be because, even if missing a chance to answer sometimes breaks a subject's concentration, the timeout message itself introduces an extra 300 ms between trials to recover.

4.3.2 Transitions

Futher analysis examined the increase in sensivity from adding CV and VC transitions back into the truncated stimuli from the preceding section. These data are relevant to the experimental hypotheses concerning the difference between /p/ and /sp/ and syllable-position asymmetries for liquids and /n/. In each of these cases, we predicted that one type of stimulus should have an advantage over the other pertaining to consonant steady-states, transitions, or both.

Examining the zero-gate stimuli, we found some evidence for the expected difference between /p/ and /sp/: all three pairs of vowel examined displayed the expected pattern; it was statistically significant for two of them. For liquids in onset and coda, one comparison came out in the expected direction: sensitivity to stimuli with liquids is much higher in coda than in onset position for [e-o], both in absolute terms and as compared to obstruents in the two conditions. The other comparison came out in the opposite direction: sensitivity is higher in *onset* than coda position for [ɑ-u]. For /n/ in onset and coda position, none of the predicted asymmetries were observed: relative sensitivity between /n/ and singleton obstruents was the same in all conditions except [e-o] onset. In that condition, sensitivity to stimuli with /n/ was significantly smaller than to stimuli with /p/.

This means that several of our hypotheses will need to be confirmed from transition data, as they were not confirmed from steady-state data. Shown below is what we have confirmed so far.

Contrast	Confirmed at gate 0?	Comments
/p/-/sp/	yes	as expected
/n/-obstruent, onset vs. coda	no	1 of 2 no effect, 1 opposite
liquid-obstruent, onset vs. coda	partially	1 of 2 as expected, 1 opposite

Table 4.1. *Summary of results so far, showing what remains to be explained by transition data.*

At the very least, then, we hope to find further evidence for the liquid and nasal asymmetries in the transition data. Patterns of sensitivity increase across gates for /n/, /sp/, and singleton obstruents are shown below; liquids will be discussed later in this section. Each series of lines shows sensitivity increasing from gate 0 to gate 1 and from gate 1 to gate 2. Note that the data from gate 1 are not used in the statistical analysis, because they include a proper subset of the acoustic material in gate 2 stimuli, and we're mainly interested in the total boost to sensitivity across the two gates. We do present these data below, however, to confirm that the the identification function increases as more acoustic material is added in, and to clarify the shape of that function.

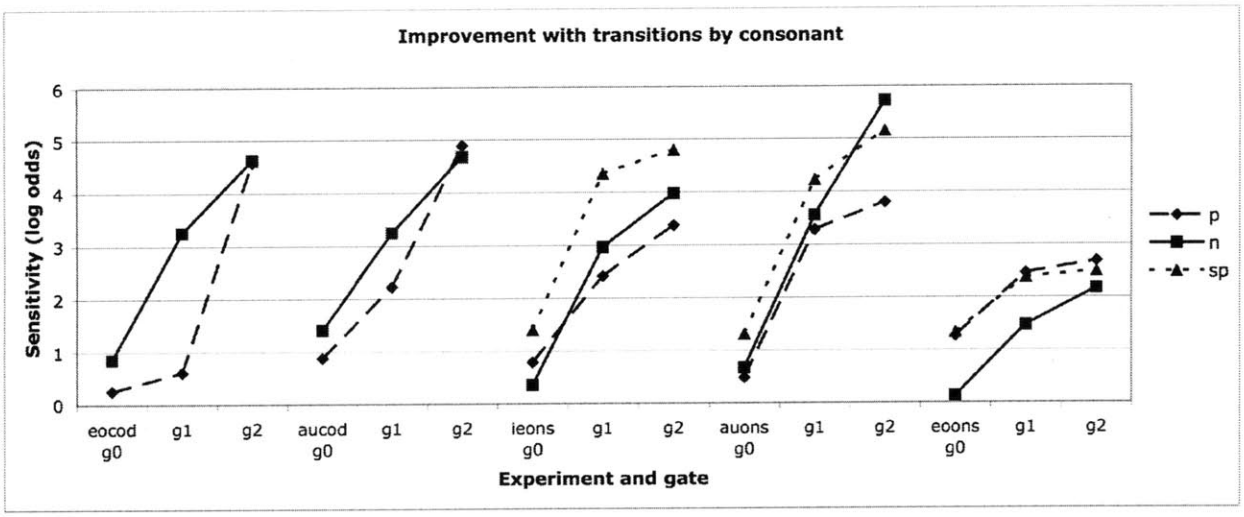

Figure 4.5. *Sensitivity by consonant, condition, and gate. The vertical axis shows the sensitivity parameter fit to each contrast by the model, in terms of differences in the log odds of a given response across stimulus categories. The name of each condition consists of the two vowels tested in the condition followed by 'ons' for consonants in onset position or 'cod' for coda.*

There are several things to notice about these data. Nasals show a larger transition increment in sensitivity than /p/ does in at least two of the three onset conditions, with a very small effect in this direction in the third. In coda, however, the increment in sensitivity associated with nasals appears to be slightly smaller than that for singleton obstruents. Differences between /p/ and /sp/ are smaller and not so easy to see here, but /sp/ appears to be associated with a larger increment in two of the three onset conditions.

A slightly more abstract way of viewing the data, which will be more germane to the statistical models described here, examines the magnitude of the transition increment associated with each type of stimulus in each condition. That data is shown below, with liquid data now added in; the figures here reflect the total increase in sensitivity from gate zero to gate two.

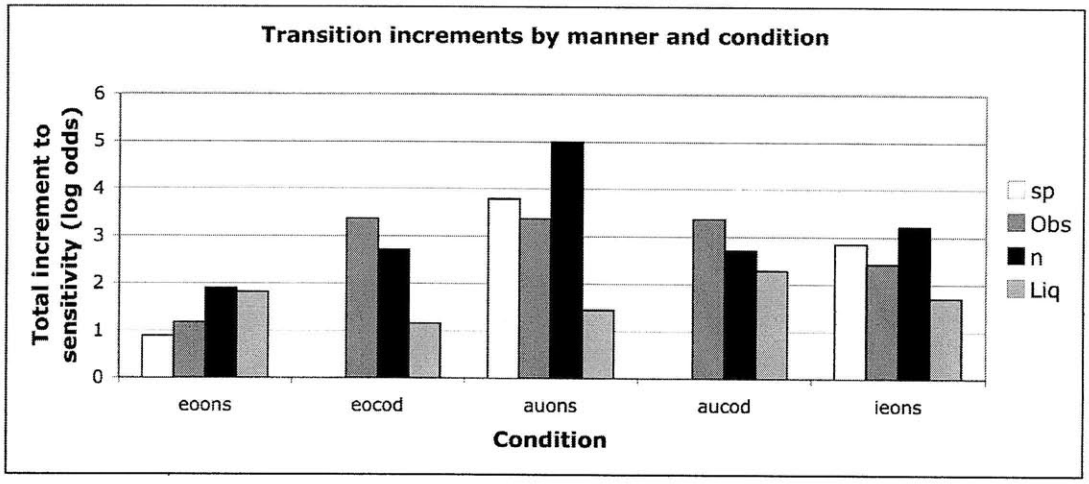

Figure 4.6. *Transition increment to sensitivity by consonant and condition. The vertical axis shows the change from gate 0 to gate 2 in sensitivity parameters fit to each contrast by the model, in terms of differences in the log odds of a given response across stimulus categories. The name of each condition consists of the two vowels tested in the condition followed by 'ons' for consonants in onset position or 'cod' for coda.*

The effects mentioned above for /sp/ and /n/ are slightly easier to see in this presentation. We can also ask about the asymmetries predicted for liquids. The prediction, recall, was that liquids should have a greater increment relative to singleton obstruents in coda than in onset position. This is true for the [ɑ-u] conditions: the obstruent transitions are much more informative than the liquid ones in onset, but the effect is smaller in coda. For the [e-o] conditions, however, the pattern is opposite.

The final model collapsed very few fixed-effect parameters for transition increments; partly, this is because there were a number of interactions across conditions that were almost but not quite significant; we retained them in case by-subject adjustments changed the picture. With regard to

transition increments, the coda nasals pattern together; the singleton obstruents in [ɑ-u] onset, [ɑ-u] coda, and [e-o] coda conditions pattern together; and the /sp/ items in [ɑ-u] onset and [i-e] onset conditions pattern together. All other contrasts were retained, although not all of them were significant in the final model.

The reversals between /n/ and singleton obstruents noted above are significant. In [ɑ-u] onset condition, nasals have a significantly larger transition increment than /p/: $\beta = 1.62$, $\chi^2 = 15.3$, $p < 0.01$. In [e-o] onset condition, the effect is smaller, but not significantly so. In both [ɑ-u] and [e-o] coda conditions, the pattern is reversed: obstruents show a larger transition increment than nasals. This reversal gives rise to a significant four-way interaction between sensitivity, gate, consonant quality, and coda position: $\beta = -2.28$, $\chi^2 = 26.4$, $p < 0.01$.

As we noted above, items with /sp/ display a somewhat larger increment than items with /p/ in two conditions. Figure 4.6 shows that the pattern reverses in the third condition. Neither the differences between /sp/ and /p/ nor the reversal between conditions reach statistical significance.

For the comparison of liquid and singleton obstruent stimuli, patterns are more complicated. In [ɑ-u] onset condition, items with /l/ show a much smaller increment than items with /p/: $\beta = -1.93$, $\chi^2 = 43.4$, $p < 0.01$. This effect is much smaller in [ɑ-u] coda condition, leading to a significant four-way interaction between sensitivity, consonant quality, gate, and coda position: β

= 0.84, $\chi^2 = 5.1$, $p = 0.02$. This asymmetry is not observed in [e-o] onset and coda conditions; there, liquid stimuli show a larger increment in onset position and a smaller one in coda.

Similar to the zero-gate model, subjects vary in their overall sensitivity and are more sensitive to contrasts from the female speaker than the male one. Both effects significantly improve model fit. For by-subject random slopes: $\chi^2 = 634$, $p < 0.01$. For by-speaker random slopes: $\chi^2 = 6.1$, $p < 0.05$.

Also agreeing with the zero-gate model, subject perform significantly worse on trials following an incorrect response on the previous trial: $\beta = -0.26$, $\chi^2 = 14$, $p < 0.01$. In this model, the effect of a timeout on the preceding trial came out nearly significant: $\beta = -0.28$, $\chi^2 = 2.8$, $p = 0.09$. As with the zero-gate model, the standard error associated with the timeout effect is larger than that associated with the incorrect-answer effect, although the size of the effects is comparable.

4.4 Discussion

The findings from this study broadly support the hypotheses put forth in chapter 3 to explain the production asymmetries observed in chapter 2. Those hypotheses are repeated here:

- Steady states of liquids contain more information about an adjacent vowel than steady states of obstruents.
- In onset position, nasal steady states or transitions or both contain more vowel information than those of obstruents, but not in coda position.

- Liquid steady states or transitions or both contain more vowel information in coda position than they do in onset position.
- Formant transitions following /sp/ clusters in onset position contain more vowel information than aspiration following /p/, or /s/ and the transient of /p/ contain more vowel information than just the transient of /p/, or both.

As for the first hypothesis, it holds in four out of the five conditions examined here. In the fifth condition, [e-o] onset, /p/ steady states appear to contain more information about the following vowel than liquid steady states, *contra* our hypothesis. However, in that condition, transitions in the liquid stimuli contain more information about the following vowel than aspiration in /p/ stimuli. It is possible that this difference in transitions is large enough to overcome the effect from the steady state. It is also possible that, although liquids and their transitions offer an advantage over obstruents for height contrasts, they do not for backness contrasts. Due to the onset/coda articulatory asymmetries for /l/ discussed in chapters 2 and 3, we would expect the tongue-tip constriction to be more overlapped with the vowel in onset than in coda position. This may constrain how much the tongue body is able to move to track the backness contrast in a following vowel. With a coda /l/ or a /p/ in any position, this constraint would not hold. In this case, the availability of incremental CS for onset clusters with liquids would be a consequence of the fact that liquids contain more information about the height of the following vowel, though not necessarily its backness.

The second prediction, concerning /n/, is confirmed in the transition data. For steady states, we found no significant differences between /n/ and singleton obstruents except for [e-o] onset

condition, where the effect went in the opposite direction from what we predicted. In the transition data, however, transitions between nasals and vowels contained more information about the vowel than singleton obstruent transitions in onset position, and less in coda position. This held for both the [e-o] and [ɑ-u] contrasts. The [i-e] contrast also had a large transition-driven increment in sensitivity for stimuli with /n/, although we have no coda data to compare it to. Again, the unexpected steady-state result in [e-o] onset condition may have to do with the articulatory asymmetries between onset and coda /n/ discussed in the preceding chapters, which are broadly similar to those for /l/. And again, it is possible though not certain that the increased information in n-V transitions relative to aspiration is enough to make up for this difference.

The third prediction, concerning asymmetries between onset and coda liquids, is confirmed partially in the steady-state data and partially in the transition data. We predicted that the transition increment in vowel sensitivity associated with liquid stimuli was larger relative to that for singleton obstruents in coda than in onset position. This was true for [ɑ-u] but not [e-o]. Because liquid steady states appear to contain less information about the [ɑ-u] contrast in coda than in onset position, relative to singleton obstruents, this is another case where the transition effect would need to overcome a steady state effect. For [e-o], the transition effect is the opposite of what we predicted. Note, however, that there was a massive asymmetry in the predicted direction for steady states. In [e-o] coda condition, the difference between liquid and /p/ steady states was the largest observed in the experiment; while in onset, the effect was reversed. The evidence that compression asymmetries in onset and coda position for liquids correspond to

perceptual patterns, then, is mixed. Although we have some suggestive results, there are also two patterns that seem to go against the hypothesis.

The final hypothesis, concerning /p/ and /sp/, is confirmed from steady-state data. Stimuli with /sp/ contained more information about the following vowel than stimuli with /p/ in all three onset conditions; the difference was significant in two of them. Transitions adjacent to /sp/ also were slightly more informative in two of three conditions, but the effect did not reach statistical significance. The 'odd condition out' for both these effects is the [e-o] onset condition.

Although all of the predictions were confirmed to some extent, almost all of them also ran into trouble with the [e-o] onset condition. For some of these, we mentioned plausible hypotheses about why this might be so. But it is also possible that there was something strange about this condition. Subjects in this condition performed far worse overall than any of the other conditions. This is despite the fact that [e-o] is one of the easier vowel contrasts to discriminate, according to the other experiments reanalyzed here. This may indicate that there was something exceptional about the subjects or the stimuli in this experiment, or both. On the other hand, the results may simply indicate that, while the predictions of the theoretical model from chapter 3 hold in general, they do not hold for every single vowel contrast.

4.5 Conclusion

The experiment described in this chapter found that many of the production asymmetries in compression discovered in chapter 2 mirror perceptual asymmetries. The relationship between

the two types of data is predicted by the duration grammar developed in chapter 3. That grammar predicted that vowels shorten more when there is more information about them in the surrounding context.

In chapter 2, we argued that compression effects are more amenable to an auditory explanation than an articulatory one. Until now, that argument was largely a negative one: there is no clear way to explain compression asymmetries in terms of gestural coordination, so we should seek other alternatives. Now, however, we've shown that there is a timing grammar based on auditory representations that can predict the attested patterns of compression; that that grammar requires certain perceptual asymmetries to hold in order to predict the attested production patterns; and that many of those perceptual asymmetries do, in fact, hold. The argument for an auditory account of compression and compensatory shortening is now considerably strengthened.

Most of this thesis so far has been concerned with small phonetic differences between English utterances. Very little has been said so far about other languages and about how the timing grammar interacts with the system of phonological contrasts, if it does at all. In the next chapter, we lay out a framework for how timing constraints may interact with constraints on phonological contrasts. Cross-linguistic examples will highlight cases where timing may interact with phonotactic licensing, and we'll see that the weighted constraint formalism is capable of analyzing each of these cases.

Appendix 4A

4A.1 Reanalysis of Whalen 1983

Whalen (1983) tested vowel identification using only post-vocalic fricative noise. Stimuli crossed the vowels {i, u, o, a} with the consonants {s, z, ʃ, ʒ} in coda position. Subjects identify the vowel correctly significantly more often than chance. Using chi-square tests on contingency tables with data pooled across subjects, he reports that subjects are above chance at discriminating the rounding contrast and the height contrast.

Count data were reconstructed from the description of the experiment, the conditional probability table (Table 2) in the paper, and the contingency tables (Tables 4 and 5) by feature and consonant in the paper. The count data were analyzed with several log-linear models. All models included terms for each stimulus (eight), and bias terms for each vowel response ({a,i,o,u}) in each consonant context ({s, ʃ}). Cell counts were the dependent variable in each model; different phenomena were tested with different cell-grouping factors.

For stimulus-response identity ('subjects identify the vowel significantly above chance level'), the grouping factor was simply whether subjects provided the correct response vowel, i.e., the original vowel that the stimulus was adjacent to. For height and roundness, the grouping factors were whether the subject responded with a vowel that had the correct height and roundness features, respectively. For bias, individual terms were compared to each other by dropping factors from the model and performing a likelihood ratio test.

Subjects identify the vowel correctly from only post-vocalic fricatives significantly more often than chance: : $\chi^2 = 26$ on 1 Df; $p < 0.0001$. Whalen reports that subjects are above chance for /i/ and /u/ stimuli, but not for /a/ and /o/ stimuli. The reanalysis shows that they are significantly above chance for /i/ and /u/ stimuli, marginally significant for /a/ stimuli, non-significant for /o/. The Bonferroni-adjusted α criterion is 0.0125. Accuracy for /i/: $\chi^2 = 51$ on 1 Df; $p < 0.0001$. /u/: $\chi^2 = 9$ on 1 Df; $p = 0.0022$. /a/: $\chi^2 = 4$ on 1 Df; $p = 0.0351$. /o/: $\chi^2 = 2$ on 1 Df; $p = 0.1792$

Whalen reports that subjects respond with vowels that have the correct height specification significantly more often than chance; he makes the same claim for rounding. In other words, matching the stimulus for rounding and for height each independently make responses more likely. In the paper, this analysis is conducted with separate chi-square tests on contingency tables for each contrast, as well as each consonant. This entails separate tests on how sensitivity and bias are affected by roundness, by height, by the interaction of roundness with consonant, and by the interaction of height with consonant. If any of these factors are correlated (and they almost certainly are), a chi-square test may fail to give an accurate picture of the independent significance of each of the effects. Also, these comparisons neglect to consider the interaction of roundness and height.

The reanalysis shows that height and roundness do interact. To put it slightly differently, once we separate the effect of getting the vowel completely correct, the independent effects of getting roundness and height correct become smaller. The effect of roundness is significant: $\chi^2 = 7$ on 1 Df; $p = 0.0098$. The effect of height is non-significant: $\chi^2 = 0.09$ on 1 Df; $p = 0.7605$. The effect of the interaction (equivalent to complete identity) is significant: $\chi^2 = 8$ on 1 Df; $p = 0.0036$.

What this means is that, considering just those responses that are not completely identical to the stimulus (i.e., incorrect answers), subjects are not more likely than chance to get height features correct.

These results are consistent with subjects being more accurate for roundness than height. To test this, I compared a model that collapses the two contrasts to one that includes both. The reduced model includes a single variable that is marked 1 when a subject gets height or rounding correct, 0 otherwise. This test shows that the difference between the contribution of the two features is significant: $\chi^2 = 8$ on 1 Df; $p = 0.0045$.

Response bias differs depending on the following consonant. In the /s/ condition, response bias follows the scale a < o < u < i. In the /ʃ/ condition, bias follows the scale i < a < o < u.

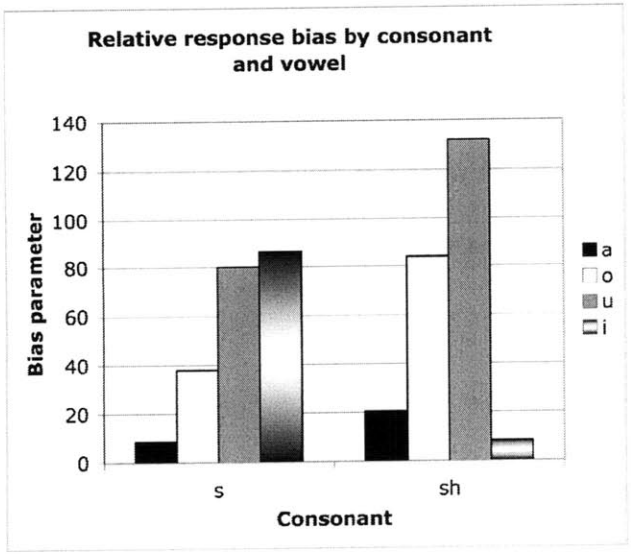

Figure 4A.1. *Response bias in Whalen's data. Groups of bars represent consonant contexts; series of bars represent vowel responses.*

Differences in bias between vowels in the /s/ condition are generally significant at two-step intervals on the scale mentioned above. For instance, the difference between /a/ and /u/ is significant: $\chi^2 = 8$ on 1 Df; $p = 0.0059$. The difference between /a/ and /o/ is not. Differences in bias between vowels in the /ʃ/ condition show a split between rounded and unrounded: both of the rounded vowels display significantly higher response bias than both of the unrounded vowels: $\chi^2 > 7$ on 1 Df; $p < 0.005$ for all comparisons. But contrasts within the rounded vowels and the unrounded vowels are not significant.

The interaction of consonant and response bias for /i/ is significant: $\chi^2 = 12$ on 1 Df; $p = 0.0007$. None of the other interactions between vowel, consonant, and response bias are significant.

There is significant sensitivity (i.e., significantly above 0, which would be chance) for all vowel contrasts except /o/-/u/. /i/-/u/ is the most distinct contrast. Figure 4A.1 shows all contrasts; on this scale, items three steps apart are significantly different in general, while items less than three steps apart are not. Sensitivity to individual vowel contrasts, to height and roundness contrasts, and in general (likelihood of correctly identifying the original vowel) does not differ between /s/ and /ʃ/ contexts.

In sum, subjects get a significant amount of vowel information from a succeeding fricative alone. They get more information about backness/roundness contrasts than they do about height contrasts. They have a significant bias to respond with higher vowels when they hear an /s/ (i, u > o > a), and with round vowels when they hear an /ʃ/ / (o, u > i, a).

4A.2 Reanalysis of Parker & Diehl 1984

Parker & Diehl (1984) examined identification from silent-center and noise-center stimuli; these are stimuli that have some central portion removed and replaced with either silence or broadband noise. The stimuli had the form /dVd/ with /i, e, u/ in one comparison set and /ɪ, ɛ, ʌ/ in the other.

Count data were reconstructed for 1 condition, 90% deletion, from a conditional probability table given in appendix B. This data concerns the vowels /ɪ /, /ɛ/, and /ʌ/. Other data is not given. The paper states in one place that there were 12 subjects, in another place that there were 16. I assumed that the figure of 16 was correct. I constructed a log-linear model with bias parameters for each vowel, a grouping variable for correct responses, and independent sensitivity variables for each contrast.

Subjects perform significantly better than chance: $\chi^2 = 18.7$ on 1 Df; $p < 0.0001$. They perform above chance for stimuli containing each of the vowels: $\chi^2 > 7$ on 1 Df; $p < 0.01$ for all factors. They show significant sensitivity to all contrasts: $\chi^2 > 40$ on 1 Df; $p > 0.0001$ for all factors. They are most sensitive to /ɪ/ - /ʌ/, less sensitive to /ɛ/ - /ʌ/, least sensitive to /ɪ / - /ɛ/. All differences are significant: $\chi^2 > 20$ on 1 Df; $p > 0.0001$ for all factors

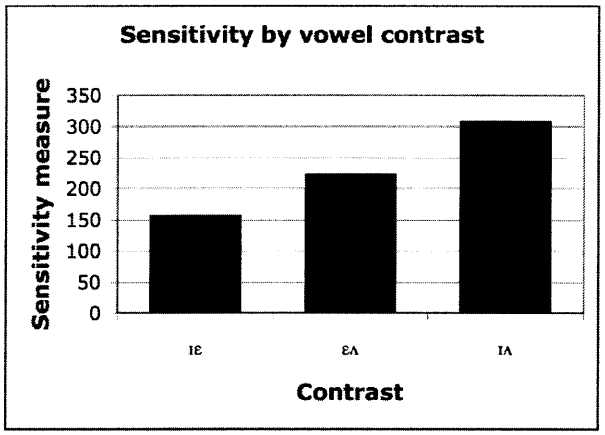

Figure 4A.2. *Sensitivity in Parker & Diehl's data.*

This is entirely consistent with the results from Whalen reanalyzed above: subjects are reasonably good at extracting vowel information from adjacent obstruents (and some transition in this case), and they recover information about backness contrasts more easily than (one-step) height contrasts.

4A.3 Reanalysis of Repp & Lin 1989

Repp & Lin (1989) investigated vowel identification from isolated transients without frication or aspiration. These stimuli may have artificially clear spectral properties due to lack of other noise components which could mask them in natural speech. They tested a wider variety of vowels than any of the other studies summarized here: consonants were {b, d, g}; vowels were {i, e, ɛ, æ, ɐ, u, o, ɔ, a}. They report that there are clear vowel spaces present in spectra of the transients, suggesting an acoustic basis for subjects' ability to distinguish vowels based on only the transients.

For the first two studies, we ran statistical analyses by fitting a loglinear model. We did not do so for this experiment. Due to the large number of items tested in this experiment, the model would have been extremely complicated to fit and interpret. And because we don't have separate count results by subject, the statistical tests from such a model would not be completely reliable anyway.

Count data were instead used to create a Biased Choice Model (henceforth *BCM*, Luce 1963). This model analyzes responses in identification experiments, distinguishing between sensitivity and bias. This model does not come with any statistical tests, but it is much more straightforward to fit and interpret than the loglinear models discussed above. The main purpose of this reanalysis was to check that the paramters of the BCM agree roughly with the other studies.

The BCM is stated as in (1):

(1)

$$p(r_j \mid s_i) = \frac{\eta_{ij} b_j}{\sum_{r_k} \eta_{ik} b_k}$$

It declares that the probability of reponse *j* given stimulus *i* ($p(r_j|s_i)$) is proportional to the similarity between *i* and *j* (η_{ij}) and the bias to respond with *j* (b_j). The summation term in the denominator normalizes based on all of the possible responses. I'll notate this term *Z* in what follows, for visual and typesetting ease.

Similarity ranges from 0 to 1 and is symmetrical. The similarity between any item and itself is 1. Given our count data, we already have the bias terms (which we equate with response frequency) and the conditional probabilities. This means that we can solve for the similarity term η:

$$\eta_{ij} = \frac{Z \cdot p(r_j|s_i)}{b_j} = \frac{Z \cdot p(r_i|s_j)}{b_i}$$

And because the distance between an item and itself is 1:

$$p(r_i|s_i) = \frac{b_i}{Z}$$

We use the second equivalence to factor the bias terms and *Z* out of the first, allowing us to state similarity measures in terms of conditional probabilities:

$$\eta_{ij} = \text{sqrt}\left[\frac{p(r_j|s_i) \cdot p(r_i|s_j)}{p(r_i|s_i) \cdot p(r_j|s_j)}\right]$$

This gives us a measure of similarity with bias factored out. Distance or sensitivity will be defined as the negative natural logarithm of η, a metric referred to as *d*.

The BCM shows that bias is generally highest for /ɛ/ and lowest for /e/, but patterns change by consonant.

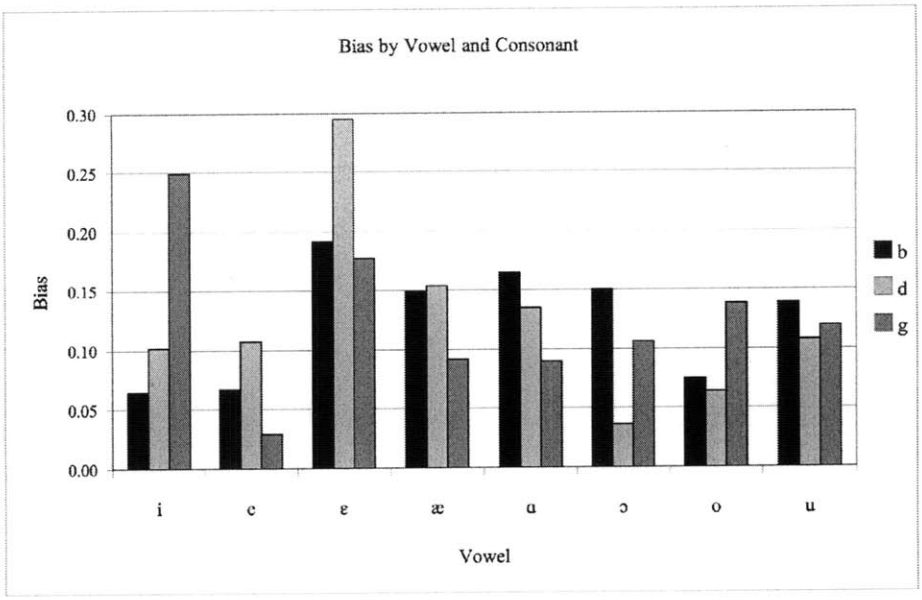

Figure 4A.3. *Bias in Repp & Lin's data.*

Senstivity similarly differs by consonant. For /b/ and /d/, height contrasts in the front of the vowel space are more distinct than height contrasts in the back of the vowel space. For /g/, height contrasts in the low part of the vowel space are much worse than height contrasts elsewhere; this is presumably because of the bottom of the vowel space is compressed with /g/, as shown by the acoustic analysis in the paper. Height contrasts in general seem to be more distinct with /b/ and /d/ than with /g/.

Backness contrasts, generally speaking, get less distinct as you go lower in the vowel space. This is not surprising given the acoustic dimensions of the space. This effect is clearest for /g/, which has extra compression low in the space. It looks like /b/ might lead to slightly more distinct backness contrasts than /d/ does, with the exception of /i/-/u/; this may have to do with compression of the back of the space with /d/.

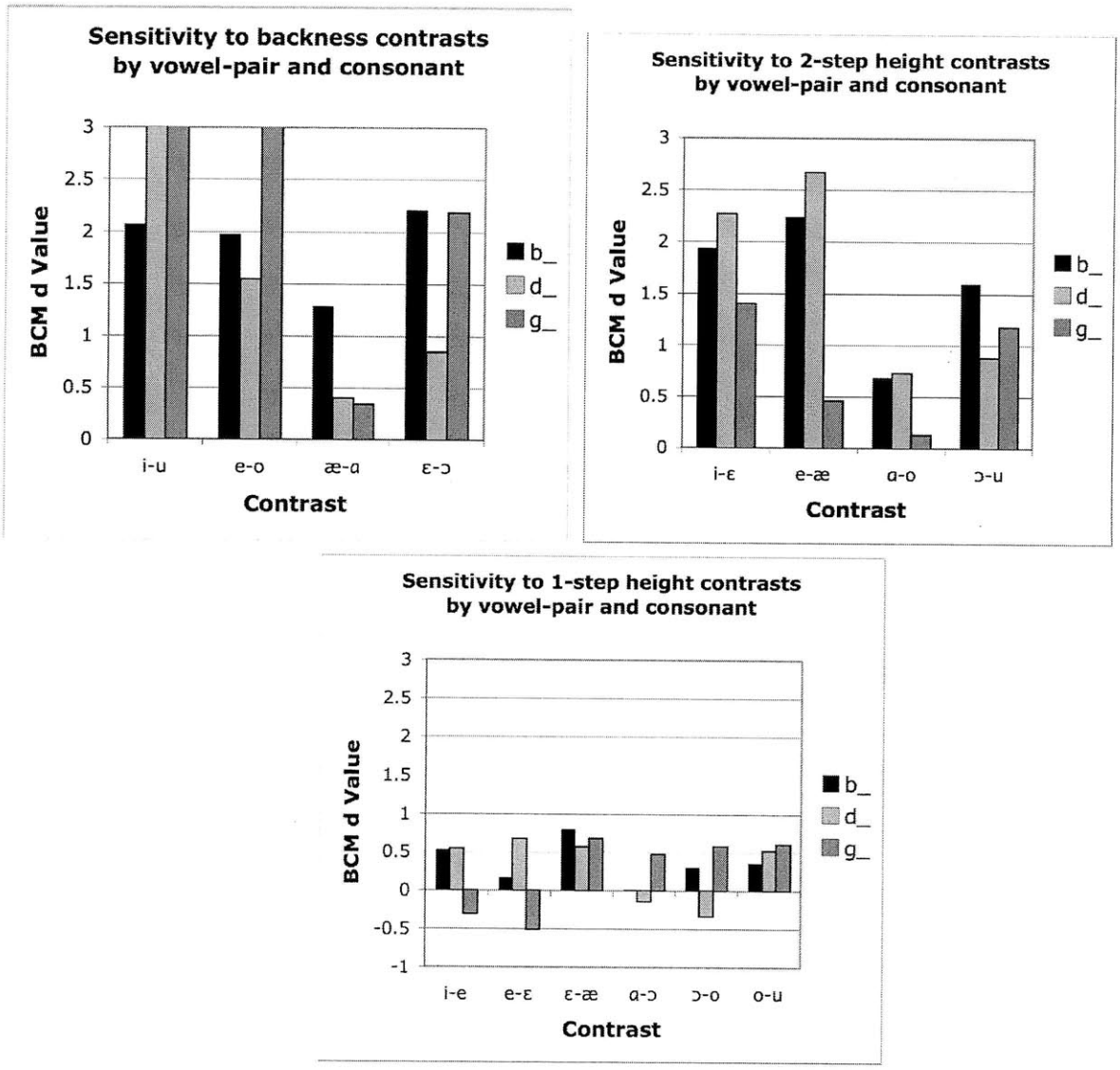

Figure 4A.4. *BCM d parameters for Repp & Lin's data. Each separate graph shows a different type of contrast: one-step height, backness/rounding, and two-step height. Negative d values indicate below-chance performance.*

The d parameters show that subjects are best at backness contrasts, less good at two-step height contrasts (e.g. /i/ vs. /ɛ/), and worst at one-step height contrasts (e.g. /i/ vs. /e/). The other experiments reviewed here consistently show that backness/rounding is easier to recover than

height, but they compare backness/rounding contrasts to one-step height. The graphs below show that two-step height is closer to backness in distinctiveness. This is not terribly surprising; contrasts like /i/-/u/ make use of the entire back/round dimension, while contrasts like /i/-/e/ make use of only a small portion of the height dimension.

The advantage for backness contrasts goes away when you consider pairs lower in the vowel space, such as /æ/-/a/. There's no way to tell whether this is because there's no lip-rounding at issue here or because the vowel space is compressed in the F2 dimension for low vowels, but these are very similar explanations anyway.

5 Timing and phonotactics

5.1 Introduction

The preceding chapters have been largely concerned with fine-grained phonetic details of English. This chapter explores how timing patterns interact with systems of phonological contrast in English and other languages. We develop a general theory of how timing interacts with phonological contrast, implemented in the weighted-constraint formalism introduced earlier. We then illustrate how this system preserves the insights of previous phonetically-based approaches to phonology, and review some novel predictions that come out of the current approach. These predictions mainly concern long-distance interactions between phonological objects; some of them appear to be corect, while others are more difficult to confirm. In the final section, we discuss the challenges that the typological facts pose for the theory of timing and phonotactics, and explore ways that the formalism might be constrained if the predictions of the current approach can not be confirmed.

Timing patterns within and between speech sounds are relevant to many aspects of phonology. Because constraints governing phonological contrasts make reference to the perceptual distinctiveness of those contrasts, and the distinctiveness of a contrast sometimes depends on the details of the timing grammar developed earlier, we predict that timing is relevant to the licensing of contrasts. This chapter focuses on how timing affects the availability of cues to phonological contrasts involving singleton consonants and clusters. By focusing on one small set of phenomena in phonology, we hope to illustrate in detail how the grammar of timing might be incorporated into phonological theory more generally.

The line of research pursued here is part of a more general program in phonology (sometimes called *licensing-by-cue*) that attempts to explain the availability of phonological contrasts partially in terms of the perceptual distinctiveness of those contrasts (Flemming 1995, Jun 1995, Silverman 1995, Steriade 1997, Kirchner 1998, Hayes *et al.* (eds.) 2004). This theory of phonology holds that patterns of phonological contrast and neutralization are determined in part by the presence of auditory cues that help listeners to tell segments apart. We briefly introduce the main ideas of this approach in the next section.

5.2 A phonetic approach to phonology

5.2.1 Licensing by cue

One of the most comprehensive analyses in this tradition is Steriade's (1997) account of the licensing of laryngeal contrasts cross-linguistically. The availability of laryngeal contrasts appears to be subject to a universal implicational hierarchy. One step on the hierarchy consists of the fact that every language which allows voicing contrasts in positions not immediately preceding a sonorant segment (e.g. Khasi, Georgian) also allows that contrast in positions that do precede a sonorant segment. In contrast, languages that allow voicing contrasts before sonorant segments may or may not (e.g. Lithuanian, Russian) allow the same contrasts before non-sonorant segments; in the latter context, the voicing contrast is neutralized.[1]

Steriade's explanation of these facts appeals to differences in the availability of auditory cues to voicing in different contexts. Specifically, voice onset time (VOT) is an important cue to stop

[1] The full implicational hierarchy has more levels. I use this single asymmetry as an illustration.

voicing contrasts for stops; this cue can be exploited before sonorant segments, but not before non-sonorant ones. The grammars of individual speakers reflect these speakers' knowledge about the perceptual distinctiveness of the voicing contrast in various positions; because grammars refer directly to distinctiveness or perceptual distance, they may favor more distinct contrasts over less distinct ones. The task of learning a particular grammar, then, consists only in finding the 'line' that your language draws to determine how distinct a voicing contrast must be to avoid neutralization.

The same approach can be extended to other phonological phenomena. In particular, the problem of *sonority sequencing* can also be analyzed with reference to the availability of cues in various contexts. Sonority sequencing refers to the tendency for every syllable to contain exactly one sonority peak, a segment more sonorous than any of the segments adjacent to it.[2] One common statement of the sonority scale is as in (1).

(1) Sonority scale

stops < fricatives < nasals < liquids < glides < vowels

Syllables that conform to sonority sequencing are syllables with a single peak; other segments in the syllable will decline in sonority as we move away from that peak segment. Sequences such as /klerp/, /kle/, and /erp/ conform to sonority sequencing. Sequences such as /lkerp/ do not,

[2] The exact acoustic correlates of sonority are controversial, but this is not such a concern in the account sketched here, where the notion of sonority is emergent from phonotactic patterns which are themselves driven by cue availability.

because both /l/ and /e/ are local sonority peaks. This chapter will focus mainly on stops and liquids.

Wright (2004) outlines how certain phonological strings that obey sonority sequencing principles will tend to maximize the perceptual robustness of phonological contrasts, especially those involving stops. One of the reasons for this is that many cues to the presence and features of stops are contained in the transitions to more sonorous segments (e.g. liquids). In the licensing-by-cue approach, contrasts are preferentially preserved in contexts with more cues or with more robust cues. This entails, among other things, that stops should be allowed to participate in more contrasts adjacent to relatively sonorous segments, and fewer contrasts adjacent to relatively less sonorous segments. The limiting case of 'fewer contrasts' is no contrast at all: in contexts with poor cue recoverability, the stop may not even be licensed to participate in a contrast between its own presence and absence. In other words, stops would not be allowed to appear in some contexts (or, depending on articulatory constraints, might be forced to appear). These would be contexts with relatively few or relatively weak perceptual cues compared to contexts where stop contrasts are licensed.

The presence or absence of auditory cues is itself partially determined by timing patterns. For instance, a burst is a useful cue to the presence and features of a stop. If that stop is tightly overlapped with a following stop, the burst may not be audible; this reduces the number of cues available. If instead the two stops are produced with little or no overlap, both bursts will be audible; this increases the number of cues available. So we don't fully know which cues will be present in a given phonological string until we know the temporal qualities of that string.

Work in the licensing-by-cue tradition has generally represented the relationship between timing and cues in the form of hypotheses about the articulatory gestural organization of sound sequences and some form of a preference for overlap. For instance, Steriade (1997) suggests gestural representations for the timing of laryngeal and oral constrictions, while predicting that those timing relationships could be changed in order to produce better cues to the laryngeal contrast. Gordon (2001) implements a similar analysis of laryngeal contrasts in Hupa, using a constraint that prefers overlap between adjacent segments to explain why bursts of pre-consonantal stops may be obscured. Jun (2002), in a discussion of place assimilation, assumes that the release of a pre-consonantal stop, and hence the timing of that stop, is governed by a violable constraint. An important point from these approaches is that we must consider ways that recoverability and cues could change with timing.

If we are analyzing why stop-stop clusters are not possible word onsets in English, for instance, we must consider various ways that they could be produced if they *were* possible. To see this, we begin with a schematic analysis of the difference between stop-stop and stop-liquid clusters. Such an analysis might propose that clusters of two stops are not possible word onsets in English because there are not enough cues to the presence or absence of the first stop. This first stop would have no VOT or formant transitions following it, and it might have its burst covered up by the following stop (although this is not always the case with heterosyllabic stop-stop sequences in English). Stop-liquid clusters, on the other hand, are allowed as word onsets because the stop will have an audible burst and VOT. This is shown in the diagrams below, which are idealized depictions of what spectrograms for the two sequences might look like.

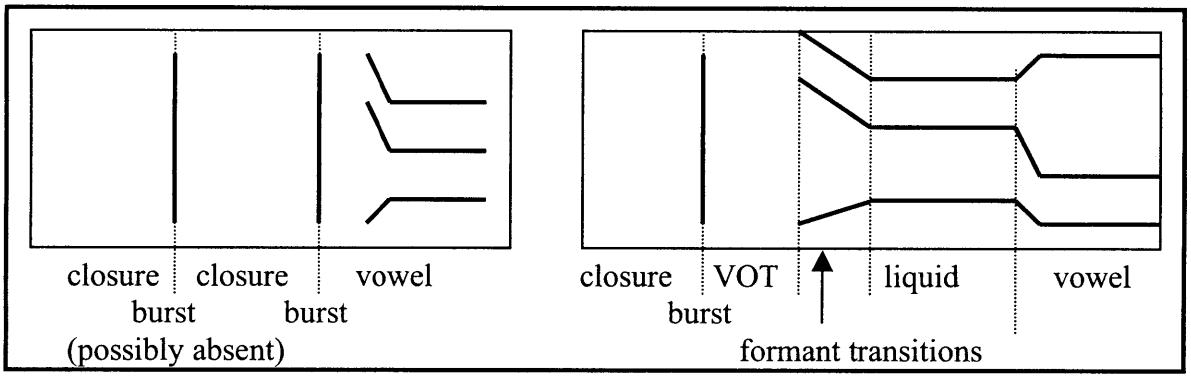

Figure 5.1. *Schematic spectrograms for a stop-stop sequence (left) and a stop-liquid sequence (right), illustrating some differences in cue availability. Note that the presence of and relationship between VOT and formant transitions depend on the voicing of the stop; the diagram on the right is meant to clearly illustrate each type of cue, but is not a likely realization for either type of stop.*

5.2.2 Temporal coordination and phonotactics

The schematic explanation above assumed that the cues available in a stop-stop sequence are known ahead of time, simply from the fact that it is a stop-stop sequence. Of course, this is an idealization: the presence or absence of cues is itself dependent on how much the two segments overlap. The analysis so far assumes that, if English had stop-stop sequences as word onsets, they would not be realized in such a way as to provide good cues to the presence of the first stop, or that there is no way to realize the first stop that will create good enough cues to license its presence. While these may be plausible hypotheses, they are not logical necessities.

We've proposed that stop-stop sequences can't serve as word onsets in English because there are insufficient cues to the presence of the first stop. We could introduce more cues, however, by

producing the two stops with an *open transition* between them. This is a period of time when the first stop has already been released, but the constriction for the second stop has not yet been formed. It would, first of all, guarantee that the release of the first stop is audible. If the stops are voiceless, the transition is most likely to take the form of aspiration, which is itself a cue to the presence of a stop, and also contains formant transitions that signal the presence and features of the stop. If the stops are voiced, the transition will most likely take the form of an *excrescent vowel* (referred to variously in the literature as *excrescent vowel/schwa*, *intrusive vowel/schwa*, *svarabhakti vowel*, and *open transition*). This is a period of sonorous, vowel-like periodic sound that contains formant structure and transitions, which will tend to increase the recoverability of a preceding stop.

In fact, this is exactly what happens in some languages. In Montana Salish, for instance, where stop-stop clusters *are* possible word onsets, they are produced with open transitions (Flemming *et al.* 2007). Shown below is a recording of a Montana Salish speaker saying the word /ttəʔwit/ 'youngest'.

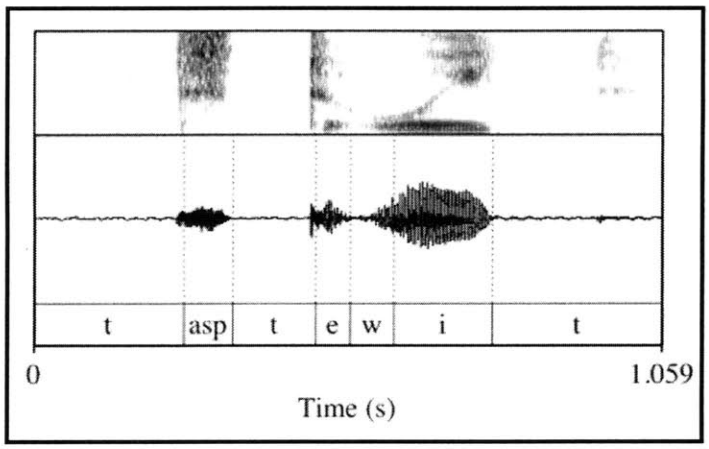

Figure 5.2. *An utterance of the word /ttəʔwit/ 'youngest' from a native Montana Salish speaker. Clearly visible in the waveform and spectrogram is an aspirated open transition between the two initial stops. Glottalization on ʔw not annotated; schwa transcribed as /e/. Audio from the UCLA Phonetics Archive.*

In between the release of the first /t/ and the closure of the second /t/ is a clearly visible (and audible) interval of aspiration. This is an open transition. It contains aperiodic noise and formant transitions, which serve as cues to the presence and features of at least the first stop.

If phonotactic licensing depends on cues, and cues depend on temporal coordination, then our toy analysis of English is missing something important. It is not enough to say that English stop-stop sequences wouldn't contain enough cues to support contrasts; we must explain why English stop-stop sequences don't surface in a way that would contain more or better cues. In other words, we must explain why English isn't Montana Salish.

One possibility is that there simply is no such explanation. In this view, temporal coordination is a parameter that varies between languages and forms part of the input to the phonotactic

grammar for an individual language. Articulatory patterns are in some sense determined 'before' phonotactic patterns, perhaps by principles of abstract gestural coordination (Goldstein *et al.* 2006, Nam *et al.* 2009). These structures in turn determine which possibilities are available to the phonological grammar. The desire for robust perceptual cues can not affect the basic properties of segmental timing, because information in the grammar only 'flows' in the other direction. If English happened to be like Montana Salish with respect to stop-stop sequence timing, perhaps /ttəʔwit/ would be a word of English.

There are several reasons why this view of the grammar can not be correct. First and foremost, perceptual information *does* affect timing patterns. The preceding chapters have argued this point at length for English. Other researchers have made similar arguments on the basis of rather different empirical phenomena. Directly relevant to the example above, Chitoran *et al.* (2002) and Wright (1996) both argue that cue preservation helps determine patterns of overlap for stop-stop sequences in Georgian and Tsou, respectively. We return to these languages later in this chapter.

A related reason for rejecting the 'articulation-first' model sketched above concerns the specific constraints introduced in chapter 4. The articulation-first model stipulates differences in stop-stop timing as an irreducible fact about English and Montana Salish. It is not related to other facts about the grammars of these languages or explainable in terms of any independently-observed facts. Designating temporal coordination as a primitive precludes any possibility of explaining it with reference to other properties of speech.

Deriving the differences between English and Montana Salish from constraints on target durations and phonotactic licensing, on the other hand, would allow us to relate timing differences to the ranking or weighting of constraints whose effects are independently observable in other parts of the grammar. Duration constraints, as we've seen in the preceding chapters, are independently needed to explain compression effects; this is true regardless of whether stop-stop clusters are possible word onsets in English, Montana Salish, or no documented language. The particular constraints on phonological contrast that we will introduce shortly are also independently needed to explain phenomena unrelated to consonant clusters; for instance, Flemming (2004) uses them to explain the prevalence of certain kinds of vowel inventories in the world's languages. [3]

In the remainder of this chapter, we integrate the duration grammar from chapter 4 with a grammar for phonological contrast. Letting the two constraint sets interact generates a range of predictions about how timing and phonotactics are related. Some of those predictions are common to any licensing-by-cue approach where the pattern of phonological contrasts can affect timing patterns; others are specific to the theory proposed here. We will examine some of those predictions and illustrate them with cross-linguistic evidence. We focus on cluster phonotactics to illustrate the point; the approach should of course apply to other areas of the grammar as well.

[3] A third possible theory would hold that phonotactic patterns are arbitrary and/or parametric, and the phonetic implementation module simply finds the best way to realize contrasts that are passed on from the abstract phonology. This theory is to be *a priori* dispreferred relative to a licensing-by-cue approach, because the latter explains phonological patterns with reference to independent facts, while the former requires an extrinsic theory of both abstract symbolic manipulations and phonetic implementation of the resultant strings of symbols. If the observed range of phonological patterns turns out to be wildly different than what phonetic motivations would predict, this abstract theory would be a natural fallback.

5.3 The unified grammar

5.3.1 Contrast maintenance and distinctiveness

To model any aspect of phonotactics will require constraints that result in categorical outcomes, i.e., stops are available in context *x* but not context *y*. All of the empirical phenomena analyzed so far, as well as the constraints associated with them, are gradient. This section describes a constraint system capable of unifying categorical and gradient phenomena. The general framework and constraints were developed by Flemming (2001, 2004), though the implementation here differs in some details.

We return to the pattern discussed in the previous section: English allows contrasts between words like /le/ and /kle/, but not between words like /te/ and */kte/. We can reason about this situation with the help of two principles. Contrasts between lexical items should preferably be very distinct, as mediated by cue availability, because it makes listeners more likely to understand what is being said. And having more contrasting sounds in a context is always preferable, because it allows us to discriminate between lexical items using fewer phonological units and hence less time. From this perspective, it would be preferable to allow /to/ and /kto/, because this offers us more possibilities for contrast; but it is also preferable for /to/ and /kto/ to be very distinct, so we don't confuse them.

Each of these principles can be formulated as a constraint. All of the constraints introduced so far are associated with individual linguistic items. Given a set of input parameters, these constraints will assign a fixed cost to any phonetic form. In order to evaluate contrasts, however, it will be

necessary to take into account more than one form at once. These constraints will assign a cost to an entire candidate inventory of contrasting items. For the current demonstration, we'll limit this inventory to subsets of the four items mentioned above.

The principle of maximizing contrasts (MaCo) assigns a benefit, in the form of negative cost, for every contrast maintained in the candidate inventory. The number of contrasts for an inventory of i items will be $(i \bullet (i - 1)) \div 2$. The cost is expressed as follows:

(2) $C_{MaCo} = -w_{MaCo} \bullet n$

Where C_{MaCo} is the cost assessed by MaCo, w_{MaCo} is the weight associated with MaCo, and n is the number of contrasts maintained in the candidate inventory.

The principle of distinctive contrasts is expressed with a minimum distance (MiDi) constraint that requires contrasts to have some minimal level of perceptual distinctiveness Δ. It will assign a cost to each contrast in the candidate inventory proportional to the amount by which that contrast falls short of Δ.

(3) $C_{MiDi} = w_{MiDi} \bullet (\Delta - D(x, y))^2$, for $D(x, y) < \Delta$

Where x and y are members of the candidate inventory, and $D(x, y)$ is the perceptual distance between them.

This cost is summed across all such (x, y) pairs in the candidate inventory. Note that the segment duration constraints proposed in chapter 3 were also described as recoverability constraints. This

means that we now have two sets of constraints with the same functional motivation, although they differ formally. We return to this issue shortly.

Given these two constraints and a theory of perceptual distance, we can assign a cost to any inventory. In the next section, we use these constraints to analyze the problem from English discussed above.

5.3.2 An illustrative analysis

To begin to analyze English, we need a theory of distinctiveness. In our earlier discussion, we equated this with the availability of cues to a given contrast, and we will continue to work with this view for the remainder of the chapter. For this simple example, let us assume that perceptual distance is more or less equal to the number of cues that signal a contrast. The numbers we will use here are mainly for purposes of illustration; they are just complex enough to illustrate the point. In an ideal analysis, we could derive perceptual distances directly from existing data on identification and discrimination; in practice, this is usually not possible.

For pairs like /to/-/kto/, distinctiveness will depend on the timing between the two stops. If the release of the /k/ is masked by the constriction of the following /t/, the contrast will not be very distinct at all. Perhaps carryover articulatory effects of the velar constriction will affect the acoustics of the /t/ burst; we assign this contrast a distance of 1. If the /k/ is audibly released, there is another cue to this contrast; its distance will then be 2. If an open transition is present

(most likely aspiration), it will make the contrast more distinct still; this contrast is assigned a distance of 3.

Pairs like /lo/-/klo/ will have more or less the same distinctiveness as a stop-stop sequence with open transition: burst and VOT/formant transitions. We therefore assign this contrast the same distance of 3. Note that this distance doesn't really depend on the timing between the two consonants; the relatively wide constriction associated with liquids will mask cues to the presence of the stop less than a following stop will (although possibly more than a following vowel). Contrasts involving the distance between a liquid and a stop (i.e., /lo/-/to/, /klo/-/kto/) will be assigned the highest perceptual distance of 4. This is because they differ in both internal and context-dependent cues.

We set Δ to 5, ensuring that even the clearest contrasts (with value 4) will still be assessed some cost. Changing this assumption would have no qualitative effect on the analyses presented in this chapter as long as the value of Δ is above the distinctiveness of some contrasts; it would predict that past some relatively high level of distinctiveness, speakers no longer make the effort to increase distinctiveness further.

We start off with the duration model developed in chapter 3, with the following parameters: w_1 (syllable constraint) = 100,000; w_2 (consonant constraint) = 1,000; t_σ (syllable target) = 32; $t_{x,y}$ (consonant target) = 13; t_z (vowel target) = 25; d_t (transition duration) = 4; j (vowel recoverability coefficient for transition) = 0.4; vowel floor is 20.4; consonant floor is 5. The

vowel-recoverability coefficient k is set to 0.1 for obstruents and 0.6 for liquids. These values predict incremental CS for liquids but not for obstruents, consistent with the English facts.

Our next task is to figure out what type of an inventory can best strike a compromise between the pressures from duration targets and the pressures from contrast maintenance. We begin by considering the context #_L, where L is a liquid. In this context, the contrast between a stop's presence and its absence will either be licensed or neutralized. In the case of neutralization, the contrast can either neutralize to the absence of a stop or to the presence of a stop. These possibilities correspond to the three inventories (labeled A-C) of two items each shown below with their associated duration-target costs (these will be referred to as *markedness costs* in what follows). These markedness costs are obtained by summing the duration constraint violations of all contrastive forms within each inventory.

(4) Candidate Inventories

Inv. A	*/k/ dur.*	*/l/ dur.*	*/e/ dur.*	*Cost (k,l,e,σ)*
le		12.75	15.25	313
kle	5.25	9.75	13	91,328
Inv. B				
le		12.75	15.25	.313
Inv. C				
kle	5.25	9.75	13	91,328

One thing to notice is that the solution labeled *C*, which involves neutralizing to /kle/, incurs a substantially higher aggregate markedness cost than candidate B, which neutralizes to /le/. This is because longer syllables are always more marked in our duration framework than shorter ones.

Because neutralizing to /le/ will always be less costly in this system than neutralizing to /kle/, and the two inventories don't differ with regard to contrast preservation, it is impossible for neutralization to /kle/ to emerge as optimal, as long as there is no other constraint that favors /kle/.[4] In the language of OT, we say that Inventory B *harmonically bounds* Inventory C with regard to this constraint set.

We can now ask how each of the inventories fares with regard to contrast maximization and perceptual distinctiveness of contrasts. The answer, of course, depends on the weights assigned to MaCo and MiDi. In terms of duration constraints, Inventory A incurs a higher cost than Inventory B. Inventory A also preserves the contrast between a stop and its absence, unlike Inventory B. If the weight associated with MaCo is larger than the sum of the cost assigned to the /le/-/kle/ contrast and the difference in markedness costs between the two inventories, then Inventory A will emerge as optimal. If this condition fails to hold, neutralization will emerge as optimal. Intuitively, these weightings correspond to situations where maximizing contrasts is either more important (contrast) or less important (neutralization) than the combined effect of duration targets and the confusability of /le/ and /kle/. One set of weights that will produce a contrast is $w_{MaCo} = 91{,}600$; $w_{MiDi} = 100$. This is shown below:

[4] There are other ways to exclude this possibility. Most plausibly, if the theory includes constraints on articulatory effort, these constraints will always prefer the singleton consonant as well. In some cases, such as excrescent stops following nasals, there might be articulatory constraints that would favor a *longer* candidate over a shorter one.

(5) Contrast preserved

Inv. A	*/k/ dur.*	*/l/ dur.*	*/e/ dur.*	*MarkCost*	*Contrasts*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
le		12.75	15.25	312.5					
kle	5.25	9.75	13	91327.5	1	-91600	3	200	240
Inv. B									
le	Neut	12.75	15.25	312.5	0	0	--	0	312.5
Inv. C									
kle	5.25	9.75	13	91327.5	0	0	--	0	91327.5

In the table above, markedness costs are followed by the number of contrasts preserved in the inventory, the benefit associated with the contrast, the perceptual distance associated with the contrast, the cost associated with that distance, and the summed cost assessed to the complete inventory. The total inventory cost is the sum of aggregrate durational markedness as described above, negative cost associated with MaCo, and the aggregate cost assigned by MiDi to all contrasts in the inventory. In (5), The high weight of MaCo compensates for the markedness costs incurred by Inventory A, allowing it to emerge as optimal. To instead derive neutralization, we simply decrease the weight of MaCo, or increase the weight of MiDi. Below are outcomes for w_{MaCo} = 10,000; w_{MiDi} = 1000, with all other parameters the same. Harmonically-bounded Inventory C will be excluded from further consideration.

(6) Contrast neutralized

Inv. A	*/k/ dur.*	*/l/ dur.*	*/e/ dur.*	*MarkCost*	*Contrasts*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
le		12.75	15.25	312.5					
kle	5.25	9.75	13	91327.5	1	-10000	3	2000	83640
Inv. B									
le		12.75	15.25	312.5					
<k>le	Neut	12.75	15.25	312.5	0	0	--	0	625

Having assessed the two candidates above, we need not consider any others. These inventories contain phonetic forms that are optimal with regard to the duration constraints. If it were possible to better satisfy the constraints on contrast by altering the temporal properties of these forms, inventories with such altered forms might emerge as optimal. But because the liquid fails to mask cues to the stop's presence, altering the temporal properties of the phonetic forms here will not change the cost assessed by the contrast constraints.[5] As such, the optimal durations for outputs will simply be those selected by the target-duration constraints.

The situation is rather different with stop-stop clusters. As we noted above, the perceptibility of the first stop will be determined in part by the amount of overlap with the following stop: the less overlap, the less the burst of the first stop will be masked by the closure of the second. We can account for this observation by positing a constraint on the audible duration of the stop's burst, including any following noise. To increase the recoverability of the stop, a speaker may produce a longer audible burst, including in some cases a full open transition. Because the phenomena addressed here concern bursts rather than closures, we treat closure duration as a constant and assume that shortening and lengthening a stop correspond to shortening and lengthening its burst. For this reason, we appeal repeatedly in what follows to the idea that lengthening a stop results in a longer audible burst; this statement reflects the simplification made here. The only part of this assumption that is crucial for the arguments here is that producing a stop with a longer audible burst requires more time than producing one with a shorter (or null) burst.

---

[5] This is something of an idealization. The liquid may mask cues to the stop's presence to some extent, and probably masks cues to its place. The general logic of the argument here will hold as long as this masking is substantially less than that induced by a following obstruent. In addition, lengthening the liquid would plausibly enhance cues to its presence and quality; we return to this issue in section 5.4.2.

We begin by examining the two candidates that correspond to those listed above for stop-liquid clusters. This is Inventory A, with contrasting CV and CCV forms that are optimal with regard to markedness costs; and Inventory B, where the contrast is neutralized to the CV item with optimal duration values. More candidates will be examined below. The optimal duration for /k/ in a /kt/ cluster in this system is 5.25, while its target duration is 13. We'll assume that a stop which is shortened by more than a third or so from its target duration is unlikely to have an audible burst when followed by another stop. For this reason, we assign this contrast a perceptual distance of 1. We carry over the weights from (5-6) above: $w_{MaCo} = 91{,}600$; $w_{MiDi} = 100$.

(7) Neutralization

Inv. A	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contrasts*	C_{MaCo}	*Dist*	C_{MiDi}	*InvCost*
te		9.5	18.5	27852.5					
k͡te	5.25	5.25	18.5	239265.6	1	-91600	3	200	175718
Inv. B									
te	Neutr.	9.5	18.5	27852.5	0	0	--	0	27853

Given these constraint weights, Inventory B, which neutralizes the /te/-/kte/ contrast, emerges as optimal. In this case, however, we *do* need to examine other candidate inventories. Specifically, lengthening the /k/ will result in more cues to its presence, driving down the cost assessed by MiDi. For concreteness, let us assume that a stop with duration target 13 must be realized with duration 8 to have an audible burst, and with duration 11 to have an open transition. This will have the effect of introducing discontinuities into an inventory's cost function as the duration of the first stop in a cluster goes up. This is because, at the durations designated above, another cue will suddenly become audible and the cost of maintaining the contrast will suddenly drop. To

illustrate this, we vary the duration of /k/ in /kte/ while holding everything else constant. The weight of MiDi is increased substantially in this chart to make the discontinuities more clearly visible.

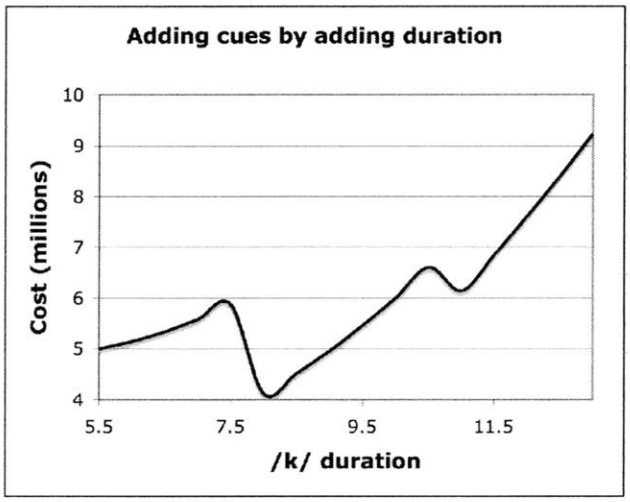

Figure 5.3. *Cost assessed to a /te/-/kte/ inventory as a function of /k/-duration in /kte/. The duration of /t/ and /e/ in /kte/ are held constant at 5.25 and 18.5, respectively. Values for /te/ are those shown in the tables above. Note the discontinuities at 8 and 11 on the x-axis, where the addition of cues to /k/'s presence reduces the cost of maintaining the contrast.*

Depending on the constraint weights, it would be possible to form an optimal inventory by lengthening the /k/ in /kte/ to preserve cues to the /te/-/kte/ contrast. This suggests the two candidates in (8).

(8) Repair strategies

Inv. C	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contrasts*	C_{MaCo}	*Dist*	C_{MiDi}	*InvCost*
te		9.5	18.5	27,853					
kte	8	5.25	18.5	1,510,453	1	-91600	2	300	1,447,006
Inv. D									
te		9.5	18.5	27,853					
k^{h}te	11	5.25	18.5	4,639,453	1	-91600	3	200	4,575,906

Inventory C lengthens the first stop so as to have an audible burst. Inventory D lengthens it even more to introduce an open transition. Note that a voiceless unaspirated stop with an audible burst is notated with no diacritic, while one followed by an open transition is notated as aspirated. Given these constraint weights, both of the candidates in (8) still incur a higher cost than Inventory B with neutralization. This is unsurprising, because MaCo is weighted very low relative to the magnitude of markedness costs. Intuititively, this means that satisfying duration targets is relatively important compared to maximizing contrasts. We discuss other weightings, which generate other languages, in the next section.

Given the constraint weights used here, then, we derive an inventory where, word-initial stops are available before liquids, but not before stops. This is the grammar of English. Note that we can also evaluate the entire inventory of liquid and stop onsets in one pass, rather than conducting separate evaluations in each consonantal context. The arithmetic changes somewhat, so the constraint weights need to change as well. The table below has $w_{MaCo} = 40{,}000$; $w_{MiDi} = 10{,}000$.

(9)

Inv. A	*/k/ dur.*	*C2 dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	C_{MaCo}	C_{MiDi}	*InvCost*
le		12.75	15.25	313				
kle	5.25	9.75	13	91,328				
te	Neutr.	9.5	18.5	27,853	3	-120,000	60,000	59,492
Inv. B								
le		12.75	15.25	313				
kle	5.25	9.75	13	91,328				
te		9.5	18.5	27,853				
kte	8.25	5.25	18.5	1,701,766	6	-240,000	170,000	1,751,258
Inv. C								
le		12.75	15.25	313				
kle	5.25	9.75	13	91,328				
te		9.5	18.5	27,853				
k^{h}te	11	5.25	18.5	4,639,453	6	-240,000	120,000	4,638,946
Inv. D								
le		12.75	15.25	313				
kle	5.25	9.75	13	91,328				
te		9.5	18.5	27,853				
k˺te	5.25	5.25	18.5	239,266	6	-240,000	240,000	358,758
Inv. E								
le		12.75	15.25	313				
te		9.5	18.5	27,853				
k˺te	5.25	5.25	18.5	239,266	3	-120,000	180,000	327,430

This table compares four-member inventories, integrating the analyses in pre-stop and pre-liquid position. Inventory A is the English-like pattern; Inventories B-D preserve the /te/-/kte/ contrast with /k/ as voiceless unaspirated, aspirated (open transition), and unreleased, respectively; Inventory E neutralizes the /le/-/kle/ contrast while preserving the /te/-/kte/ contrast. With this set of parameters, Inventory A still emerges as optimal. This is the English-like inventory. The evaluation of four-member inventories is more complicated than the two-member inventories

used above. Except for cases where the comparison of more than two forms is crucial, we continue to use the pair-wise contextual approach in what follows.

5.4 Typology and repairs

5.4.1 Selecting other stop contrasts

We've seen that the duration model developed in chapter 4, when coupled with a theory of contrast maintenance, can derive an English-like phonotactic pattern for this fragment of the grammar. The key observation is that English is prevented from 'repairing' stop-stop clusters by drastically reducing the amount of overlap that obtains within those clusters; the reason for this is the syllable-duration constraint introduced in chapter 3. On the other hand, when a cluster can be produced with relatively good cues in the presence of substantial overlap, such as /#kl/, it is possible to strike a compromise between duration constraints and the desire to maximize contrasts. The prediction, then, is that less distinct contrasts between singletons and clusters (e.g. /te/-/kte/) will never be licensed in the grammar when more distinct contrasts (e.g. /le/-/kle/) are neutralized; this is a general prediction of any cue-based approach. The grammar developed here goes a step further, explaining in terms of independently-observed constraints why some deficient contrasts in English can not be turned into more distinct contrasts by altering their temporal properties.

This property of English is a consequence of the particular constraint weights used in the analysis above. If we change the relative weights of the constraints, we predict a range of different phenomena. In this section, we show how the approach developed here preserves the

insights of the licensing-by-cue approach, while allowing for a unified account of phonological contrast and non-contrastive duration phenomena. As a first step towards deriving different patterns of contrast, consider again the four candidate inventories for the presence or absence of a stop in pre-stop position.

(10)

Inv. A	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	*Dist.*
te		9.5	18.5	27852.5		
k˺te	5.25	5.25	18.5	239265.6	1	1
Inv. B						
te	Neutr.	9.5	18.5	27852.5	0	--
Inv. C						
te		9.5	18.5	27852.5		
kte	8	5.25	18.5	1510453	1	2
Inv. D						
te		9.5	18.5	27852.5		
k^hte	11	5.25	18.5	4639453.1	1	3

Each candidate is associated with an inherent (given these constraint weights) durational markedness cost assessed by duration constraints. MaCo and MiDi will introduce additional costs or benefits based on the contrasts preserved in each inventory. Whether the effects of MaCo and MiDi are enough to overcome differences in inherent markedness between the inventories is entirely a function of constraint weighting. To favor more marked candidates (A, C, and D) that preserve contrasts will require a high weight for MaCo. To favor candidates such as D that incur high markedness costs in order to better preserve cues to contrast will require a high weight for MiDi as well. Under the hypothesis that differences between languages consist

partially or completely of differences in constraint weights, each of these possibilities should be a possible grammar.

To illustrate, we hold the markedness costs constant and adjust the weights of the contrast constraints. First, we set both contrast constraints relatively high: w_{MaCo} = 7,420,000; w_{MiDi} = 700,000.

(11) Open Transitions

Inv. A	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	***InvCost***
te	27852.5					
k$^{\text{h}}$te	239265.6	1	-7420000	1	11200000	4047118
Inv. B						
te	27852.5	0	0	5	0	55705
Inv. C						
te	27852.5					
kte	1510453	1	-7420000	2	6300000	418306
Inv. D						
te	27852.5					
k$^{\text{h}}$te	4639453.1	1	-7420000	3	2800000	47306

Inventory D, which incurs a relatively large cost from duration constraints but creates a very distinct contrast through open transitions, emerges as optimal under these weights. This grammar corresponds to languages such as Montana Salish (example above) and Georgian (Chitoran 1998), which typically include an interval of aspiration in word-initial stop-stop sequences.[6]

[6] Note that in Montana Salish, stop-liquid clusters are also repaired, by the insertion of a schwa. This may pertain to the realization of the glottalization contrast on liquids (Flemming *et al.*

By lowering the weight of MaCo and MiDi, we can select either of the other two repaired stop-stop sequences as optimal. This is shown respectively in (12-13) for w_{MaCo} = 4,200,000; w_{MiDi} = 300,000, which selects the voiceless unaspirated /kte/; and w_{MaCo} = 1,820,000; w_{MiDi} = 100,000, which selects the unreleased /kˀte/.

(12) Voiceless unaspirated

Inv. A	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	***InvCost***
te	27852.5					
kˀte	239265.6	1	-4200000	1	4800000	867118
Inv. B						
te	27852.5	0	0	5	0	55705
Inv. C						
te	27852.5					
kte	1510453	1	-4200000	2	2700000	38306
Inv. D						
te	27852.5					
k^hte	4639453.1	1	-4200000	3	1200000	1667306

2007). It also may aid the perception of a preceding stop, because these liquids tend to be pre-stopped themselves.

(13) Unreleased

Inv. A	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	***InvCost***
te	27852.5					
k˺te	239265.6	1	-1820000	1	1600000	47118
Inv. B						
te	27852.5	0	0	5	0	55705
Inv. C						
te	27852.5					
kte	1510453	1	-1820000	2	900000	618306
Inv. D						
te	27852.5					
k^hte	4639453.1	1	-1820000	3	400000	3247306

The pattern in (12) corresponds to languages such as Tsou (Wright 1996), which generally realize word-initial stop-stop sequences with an audible burst for the first stop, but no open transition. Wright reports that the vast majority of word-initial stop-stop clusters conform to this description: "In approximately 92% of the cases… the release burst alone was evident in the signal, but in a few cases the release was accompanied by a brief period of low-amplitude aspiration." (p. 76)

The pattern in (13) is more problematic. There are few or no reports in the literature of languages that allow word-initial stop-stop clusters with the first stop unreleased. The apparent rarity of this pattern may show that the cues here have been inaccurately characterized. We assigned this contrast a perceptual distance of 1, fitting with our simplified, discrete characterization of distinctiveness. In reality, the distinctiveness of this contrast may be closer to 0. As noted above, the only conceivable acoustic difference between the two items in Inventory A would be a slight

coarticulatory effect of the C1 constriction on the burst of C2. This effect may be so small as to license a contrast only at extreme weights of MaCo relative to MiDi; the contrast may be nearly impossible to sustain in a grammar.

Finally, we can predict a grammar where the stop-∅ contrast is neutralized before both obstruents and liquids. This language doesn't allow clusters with liquids and obstruents at all. The parameters $w_{MaCo} = 10{,}000$; $w_{MiDi} = 10{,}000$ will derive this system.

(14) Neutralization everywhere

Inv. A	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	***InvCost***
te	27852.5					
kʔte	239265.6	1	-10000	1	160000	417118
Inv. B						
te	27852.5	0	0	5	0	55705
Inv. C						
te	27852.5					
kte	1510453	1	-10000	2	90000	1618306
Inv. D						
te	27852.5					
kʰte	4639453.1	1	-10000	3	40000	4697306
Inv. E						
le	312.5					
kle	91327.5	1	-10000	3	40000	121640
Inv. F						
le	312.5	0	0	5	0	625

Inventories B and F, both with neutralization, emerge as optimal. This is a language that allows no clusters in word-initial position, such as Hawaiian (native vocabulary, Elbert & Pukui 1971) and Yawelmani (Newman 1944).

5.4.2 A note on the formalism

The typological analyses above illustrate a point mentioned earlier in this chapter but put aside until now: the constraints on segment duration/recoverability and on the distinctiveness of contrasts are in some sense playing the same role in the analysis. We hypothesized above that lengthening the burst and release of a stop, for instance, increases its perceptibility. Trying to produce a stop with a relaively long duration will do much the same thing. For vowels and liquids, which rely more on internal cues, perceptibility of their presence and features should increase as their duration increases, all else being equal. This means that the cost assessed to contrasts involving these sounds will vary with the extent to which they fulfill their target durations.

This raises the possibility that segmental duration targets may be removed from the theory and replaced with MiDi constraints. Indeed, in the analyses above, the costs assessed by segment duration targets for stops always correlate with the costs assessed by MiDi. If we considered contrasts involving liquids and vowels, the same would be true.

Below are evaluations of a few candidates that cross vowel contrasts, cluster contrasts, and durational realizations; this analysis eliminates segment duration constraints and floor durations,

using MiDi constraints in their place. For stops, MiDi is evaluated the same way as it was above. For vowels, MiDi is assessed solely with regard to duration, completely abstracting away from differences in quality. The assumption is that differences in formant values of the same size will be more perceptible with longer than with shorter vowels. We hold formant differences constant at some value x in order to make the problem more tractable, then set the minimum distance Δ for vowel contrasts equal to 25 duration units of formant difference x. This means the cost of MiDi for vowels will be evaluated exactly the same as the vowel duration targets above, except it will be assessed to vowel contrasts rather than individual vowels. The cost will be the square of the shortfall between Δ and the vowels' duration.

A few additional simplifications are made below. First, for contrasts involving the presence or absence of a stop, we only assess MiDi violations for minimal pairs with regard to this property. That is, we assess MiDi violations for [e-te] but not [e-to]. This is simply to make the problem more tractable. In addition, we make the simplification used earlier that the duration of CV transitions are set at a constant value and are not manipulated by the grammar. Constraint weights are all set to one, to show the raw cost profile of the candidates.

The columns show, left to right: the items in each inventory, the duration of each segment, the duration of the syllable, total durational markedness, the cost assessed by MaCo, the list of minimal contrasts for each inventory, the cost assessed to each minimal contrast, the total cost assessed by MiDi for consonant contrasts, and the total cost assessed by MiDi for vowel contrasts.

(15) Crossed contrasts, no segment duration constraints

Inv. A	Dur_{C1}	Dur_{C2}	Dur_V	Dur_σ	C_σ	*Mark Tot*	C_{MaCo}	*Contr.*	C_{MiDi}	C_C	C_V
o			25	25	0			o-to	4		
e			25	25	0			e-te	4		
to		10	18	32	0			to-te	19.36		
te		10	18	32	0			to- kˀto	16		
kˀto	7	7	14	32	0			te- kˀte	16		
kˀte	7	7	14	32	0	0	-15	kˀto-kˀte	75.69	40	95.05
Inv. B											
o			25	25	0			o-to	4		
e			25	25	0			e-te	4		
to		10	18	32	0			to-te	19.36		
te		10	18	32	0			to- kˀto	16		
kˀto	7	7	18	36	16			te- kˀte	16		
kˀte	7	7	18	36	16	32	-15	kˀto-kˀte	22.09	40	41.45
Inv. C											
o			25	25	0			o-to	4		
e			25	25	0			e-te	4		
to		10	18	32	0			to-te	19.36		
te		10	18	32	0			to- $k^h$to	4		
$k^h$to	11	7	14	36	16			te- k^hte	4		
k^hte	11	7	14	36	16	32	-15	$k^h$to-k^hte	75.69	16	95.05
Inv. D											
o			25	25	0			o-to	4		
e			25	25	0			e-te	4		
to		10	18	32	0			to-te	19.36		
te		10	18	32	0	0	-6			8	19.36
Inv. E											
o			25	25	0			o-to	4		
e			25	25	0			e-te	4		
to		10	22.4	36.4	19.36			to-te	0		
te		10	22.4	36.4	19.36	38.72	-6			8	0

The table evaluates five candidate inventories. Candidate A preserves all contrasts while compressing each segment so as to perfectly satisfy the syllable duration constraint. Candidates B and C also preserve all contrasts, but violate the syllable duration target by lengthening the vowels and first stops, respectively, in CCV words to enhance contrasts. Candidates D and E neutralize the CV-CCV contrast, while realizing CV syllables in such a way as to perfectly satisfy the syllable target and the MiDi for vowels, respectively.

Note that the cost of violating MiDi is on a different scale for vowel contrasts than it is for consonant contrasts. The former pertains to duration shortfalls, while the latter pertains to roughly the number of cues available. The claim is not that these two scales of distinctiveness stand in a fixed quantitative relationship; rather, they may be governed by separate MiDi constraints. Alternatively, there may be a single MiDi constraint that governs both types of contrast, with distinctiveness expressed along a single dimension such as confusability. We do not currently have experimental evidence that would allow us to assess distinctiveness across these different types of contrast, but such a theory might in principle be possible.

As can be seen in table (15), the cost assessed by MiDi constraints trades off against the costs assessed by durational markedness; this is how the system captures the duration-trading effects that originally motivated segmental duration constraints. For instance, candidate C lengthens the first stop in CCV sequences to enhance its perceptibility; it incurs low MiDi costs relative to the shorter stop in candidate A, but higher durational markedness. Which candidate emerges as optimal depends on the relative weights of the two constraints. Whether any candidate preserving

this contrast emerges as optimal depends on the weight of MaCo relative to MiDi and durational markedness.

It should be evident that the full range of compression effects can only be derived by evaluating multiple contrasts in parallel, and that evaluating multiple contrasts in parallel causes the number of candidates to explode. Only a tiny subset of the possible candidates are displayed above, to illustrate a general point about the logic of the formalism. Specifically, for a set of *n* contrasting items, there will be 2^n inventories differing in which contrasts they preserve (including the empty inventory), and an infinite number of possible durational realizations within each of those inventory types. As a practical matter, then, we will continue to use the system of segmental duration targets and to limit our attention to specific contexts in what follows.

There is also a non-trivial theoretical issue at play here. Evaluating a large number of lexical items in parallel, up to the entire lexicon, involves a huge number of computations. It is not clear how a language learner could effectively explore the space of possible grammars in such a system. Therefore, it may be necessary to limit the scope of the evaluation process in some way. We return to this issue in section 5.5.4 and discuss some possible approaches to constraining the system.

5.4.3 The interaction of compression and contrast

The simulation above was carried out entirely by adjusting the weights of the contrast constraints, while holding duration constraints constant in a configuration that produces English-

like compression effects. Other weights for the duration constraints (and targets) would, of course, be possible. In fact, only very specific manipulations of the contrast-constraint weights will result in the preservation of contrast given these specific weights for duration constraints; all other manipulations result in neutralization.

It is also possible to categorically change the output of the grammar by adjusting the duration constraints while holding the contrast constraints constant. For instance, lowering the weight of the syllable-duration constraint will have the effect of licensing more separation between the two stops; this in turn will render the repaired stop-stop clusters less costly.

The model largely predicts, then, that compression effects and cluster repairs are related in their motivation but independent in the surface inventory of a language. We could analyze a language with very strong compression effects in general (e.g. substantial shortening of all segments in /kle/) but blocking of those effects where preserving consonant duration improves a contrast (e.g. less shortening of C1 in /kte/). We could also analyze a language with almost no compression in general but neutralization of just those contrasts which would require substantial syllable-lengthening in order to be sufficiently distinct. We might predict that languages with more compression, all else being equal, are less likely to repair clusters using open transitions than languages with less compression, but this would depend on a theory of how constraint weights are distributed cross-linguistically. So at first, it appears that the approach makes very few concrete predictions about the relationship between phonological typology and compression effects.

Given that duration constraints are relevant to both types of phenomena, however, we can derive some more abstract predictions. The logic of the constraint system entails that, even when the effects of one constraint on some item are obscured by higher-weighted constraints, those effects may still emerge for other items when the higher-weighted constraints are rendered inactive for some reason. To be concrete, consider the grammar fragment in (11) above, which produces outputs as in figure 5.4.

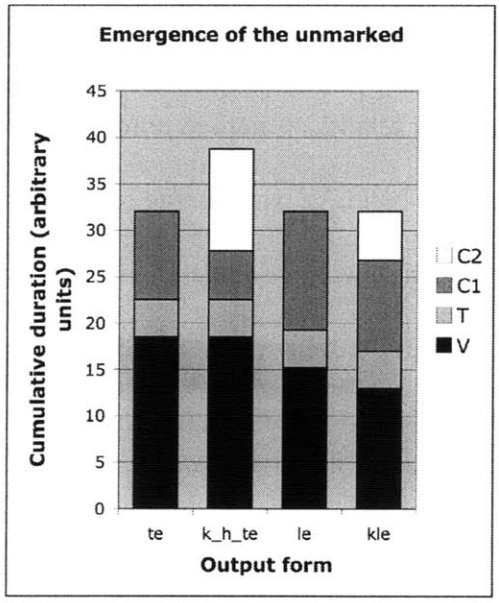

Figure 5.4. *Output forms for the constraint system in (11) above.*

In this system, /te/ contrasts with /k^{h}te/ and /le/ contrasts with /kle/. Examining the /te/-/ k^{h}te/ pair, there is substantial shortening of the /t/; but the /e/ doesn't shorten at all and the /k/ is close to its full target duration, longer even than the /t/ in /te/. With regard to C1 and the vowel, then, this sequence shows little evidence of compression. In the /le/-/kle/ pair, however, where cue availability and minimum vowel duration are both less problematic, compression emerges

strongly for all segments. This phenomenon is similar but not identical to the principle known in OT as the *emergence of the unmarked*. Here, the strong effects of the syllable-duration constraint only emerge completely when extrinsic pressures from conflicting constraints are rendered less relevant.

This property of the constraint formalism results in a concrete prediction about the relationship between compression and phonotactic licensing: we can derive patterns like that in figure 5.4, where the effects of duration constraints become more pronounced in contexts with better cues to contrasts; but we can never derive a reverse effect. This means that we shouldn't ever observe a case where items are realized with a relatively marked temporal configuration in an environment with relatively good cues to a given contrast, while being realized with a less marked temporal configuration in an environment with fewer or weaker cues to that contrast. So, for instance, we predict that no language that includes the four strings in figure 5.4 will display compression of all segments from /te/ to /kte/ but fail to display compression for some segments from /le/ to /kle/.[7]

In order to test this putative implicational universal, we need access to detailed data on the realization of clusters and the presence or absence of compensatory shortening in many languages. Such data currently does not exist, and might take years to gather. As such, we leave the confirmation of this prediction for future research.

[7] This holds unless some independent fact about cues changes the calculus. For instance, if there were a contrast that somehow becomes more distinct in the presence of overlap, the prediction would be reversed.

5.4.4 Further effects on stop-stop clusters

We saw above that, when duration costs or cue availability are less of an issue, compression effects are predicted to emerge more clearly. This prediction is borne out in several phenomena from Tsou and Georgian, which have both been mentioned repeatedly in the preceding discussion. These effects and the explanation for them were first brought to light by Wright (1996) for Tsou and Chitoran *et al.* (2002) for Georgian, but neither effect has been analyzed in a formal grammatical system. This section implements such an analysis.

5.4.4.1 Place effects

In Georgian, there are asymmetries in timing between different stop-stop clusters. Specifically, in C1-C2 clusters where the constriction of C2 is anterior to the constriction of C1 (e.g. /kt/), we observe less articulatory overlap (Chitoran *et al.* 2002) and a longer interval of time between the two acoustic bursts (Chitoran 1999) than we do in clusters where C1 is anterior to C2 (e.g. /tk/).

The explanation for this lies in masking effects. When two stops overlap extensively, the back-to-front or front-to-back order of the two successive constrictions will produce different acoustic effects. In a front-to-back sequence such as /tk/, the burst of /t/ will be audible to some extent even if the /k/ constriction is already being formed as the /t/ is released. The burst in this case presumably won't be as intense as, for instance, before a vowel, because the tongue-body constriction for /k/ is antagonistic to allowing pressure to build up behind the constriction at the alveolar ridge, but the burst shouldn't be completely masked unless the /k/ constriction is

completely blocking flow at the time of release (and even then, there might in principle be some weak acoustic reflex of the release). In back-to-front sequences, on the other hand, releasing C1 while C2 is already being formed is particularly likely to render the C1 burst inaudible. Releasing the pressure behind one constriction to allow air to flow forward into a more anterior constriction won't necessarily have any acoustic effect at all. In other words, the C1 burst is more likely to be masked in /kte/ than it is in /tke/.

For this reason, we expect that, for stop-stop sequences with a given degree of overlap, C1 will be obscured less in a front-to-back cluster than in a back-to-front cluster. It also follows that, to achieve a given level of distinctiveness for the presence of C1, front-to-back clusters will require less temporal separation than back-to-front clusters.

Chitoran *et al.* (2002) invoke this explanation for the timing asymmetries, and suggest that it would be possible to implement in recent versions of the Articulatory Phonology framework (Browman & Goldstein 2000, Goldstein *et al.* 2006). In the newer versions of this framework, consonant gestures are essentially said to repel each other when they occur in sequence. This property itself is said to emerge from recoverability considerations. The framework can express greater or lesser tendencies towards repulsion by adjusting the *bonding strengths* that obtain between particular sequences. To analyze the Georgian facts, they would say that the bonding strength governing the repulsion relationship between consonants in sequences like /kt/ is greater than that for /tk/.

The analysis adopted here is conceptually similar to this sketch, but implemented in a formalism rather different from Articulatory Phonology. Rather than specifying different bonding strengths for sequences that are easier and harder to recover with overlap, we let the overlap facts emerge from general considerations of cue availability. Because the two types of stop-stop cluster differ in the amount of time that must separate C1 release and C2 in order for C1 to be perceptible, they also differ in how long the acoustic burst of C1 must be to achieve some given level of perceptibility. Note that, in this approach, the grammar does not manipulate articulatory gestures and overlap *per se*. but rather the acoustic consequences of that overlap. So, for instance, to require that the burst from one consonant not be covered by the constriction of the next, we specify that the first consonant's burst should be realized with some non-trivial auditory duration; this latency will depend on the acoustic onset of the second consonant's constriction, thereby indirectly favoring, via acoustics, certain articulatory configurations over others.

Concretely, if the presence of a burst achieves some level of distinctiveness x when separated from the following closure by y ms in a back-to-front sequence, then a front-to-back sequence can achieve distinctiveness x with less than y ms separation. We represent this situation in our formalism by assigning cue-driven increments in perceptual distance to contrasts like /ke/-/tke/ at shorter durations for the C1 burst; contrasts like /te/-/kte/ will need longer burst durations to 'reach' the cue-driven increments associated with bursts and open transitions.

At a first pass, let us assume that bursts and open transitions emerge 'automatically' after the burst of C1 in a stop-stop cluster reaches a given duration threshold; this simplification will be discussed below. We've been working with target durations of 13 for stops; recall that we make

the simplified assumption that closure remains constant across different realizations, meaning that burst duration is manipulated by manipulating C1 duration. For front-to-back sequences like /tk/, we introduce a burst at 7.5 units of duration and an open transition at 10 units. For back-to-front sequences like /kt/, we hypothesized that to reach comparable levels of distinctness will require more time. For these sequences, we introduce the first increment at 9.5 units of duration and the second at 12 units. The table below shows candidate inventories for the context #_ke with neutralization to ∅ and /t/ durations of 7.5, 10, and 12 units. Parameters are $w_{\sigma} = 1{,}000$; $w_{seg} = 1{,}000$; $w_{MaCo} = 170{,}000$; $w_{MiDi} = 10{,}000$; all other parameters are the same as the previous analyses.

(16) Front-to-back clusters

Inv. A	*/t/ dur.*	*/k/ dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
ke	Neutr.	10.75	19.75	17943	0	0	--	0	35886
Inv. B									
ke		10.75	19.75	17943					
tke	7.5	9.5	18.5	106806	1	-170000	2	90000	44749
Inv. C									
ke		10.75	19.75	17943					
t^{h}ke	10	6.5	18.25	117063	1	-170000	3	40000	5006
Inv. D									
ke		10.75	19.75	17943					
t^{h}ke	12	5.75	18.25	138493	1	-170000	3	40000	26436

With these parameter settings, Inventory C emerges as optimal. This candidate has an open transition between the two stops. Candidate D has an open transition as well, and presumably a longer one. But it incurs needless markedness costs by lengthening the first stop past the point

where an open transition is audible. Because this system directly regulates auditory cues, the optimal solution here is to lengthen C1 just enough to maximize the cues to its presence or absence, and no more.

Compare now an evaluation of the context #_te, with the same parameter settings and candidate inventories.

(17) Back-to-front clusters

Inv. A	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
te	Neutr.	10.75	19.75	17943	0	0	--	0	35886
Inv. B									
te		10.75	19.75	17943					
k'te	7.5	9.5	18.5	106806	1	-170000	1	160000	114749
Inv. C									
te		10.75	19.75	17943					
kte	10	6.5	18.25	117063	1	-170000	2	90000	55006
Inv. D									
te		10.75	19.75	17943					
k^{h}te	12	5.75	18.25	138493	1	-170000	3	40000	26436

The solution here is still to lengthen C1 just until an open transition becomes audible. In back-to-front clusters, however, this doesn't occur until C1 is 12 units long. The result is a longer audible C1 burst, and therefore a longer interval between the onsets of the two bursts (continuing to assume that closure duration is not manipulated).[8] This is exactly what Chitoran (1999) finds for Georgian.

[8] Another possibility is that the differences in inter-burst interval reflect a longer duration for the closure of C2 in back-to-front sequences than front-to-back. In this case, the asymmetry would

5.4.3.2 Positional effects

A second asymmetry in timing for stop-stop sequences occurs in both Georgian (Chitoran 1998, Chitoran *et al.* 2002) and Tsou (Wright 1996): stop-stop sequences are more overlapped and less likely to display an audible release of C1 in intervocalic position than in word-initial position. In Tsou, for instance, C1 in word-initial stop-stop clusters is always audibly released, but only about 60% of comparable stops in word-internal clusters are released.

The reason for this is that word-internally, C1 is preceded by a vowel. The transition from the vowel into C1 closure provides valuable cues to the presence and features of a stop, so the cues associated with release become less important. There is a trading relationship between cues in the preceding transition and cues associated with the burst and succeeding transition.

To analyze this asymmetry in the current framework, we need a theory of how costs are assigned to sequences of more than one syllable, and a theory of the syllable structure of sequences such as /akta/. This is not a question with a generally-accepted answer; Chitoran *et al.* (2002) report that neither linguists nor native Georgian speakers agree on the syllabic affiliations of the consonants in such sequences. Wright (1996) notes that Tsou lacks word-final consonants, and that most but not all clusters attested word-initially are also attested word-medially. These facts could be seen as evidence that /a.kta/ is the proper syllabification, although this still leaves a residue of unexplained variation between initial and medial contexts.

---

still be a case of lengthening the overall duration of a cluster to preserve cues, but the added duration would come from the inherent minimum duration of closure for C2 rather than the release of C1.

For the purposes of this analysis, we will treat intervocalic stop-stop clusters as complex onsets. The main reason for this move is that it requires no revision to the formalism already introduced, which has been formulated with syllables in mind. We treat the first vowel as a separate syllable; presumably it fully satisfies its duration target and incurs 0 cost. The only problem, then, is to work out the markedness of the second syllable.

Note that the analysis could be modified to work with different assumptions, but these would require either more computation or revisions to the formalism. For instance, we could assume syllabification as in /ak.ta/, with the timing between the two stops (and hence the two syllables) determined by the realized duration of /k/; this wouldn't fundamentally change the analysis but would add a second markedness computation to each candidate. Alternatively, we could assume an entirely different domain for duration targets and constraints; nothing we've seen so far would distinguish between targets for syllables, feet, or prosodic words. Conceivably, there could be targets for entirely different units, such as bigrams or syllable onsets and codas. Only careful comparison of timing relations would allow us to distinguish between these possibilities, but we don't need to know which one is correct in order to analyze the timing patterns described above. The only crucial assumption is that shorter clusters in /VCCV/ sequences are less marked than longer ones.

Given our treatment of the first vowel as essentially non-existent for purposes of durational markedness, the only difference between word-initial and word-internal stop-stop clusters in our formalism will be the number of available cues to the presence and features of C1. We assign to word-initial contrasts the same values that we've been using above; word-internal cluster

contrasts will be assigned a distance increment of 1 extra unit relative to their respective word-initial counterparts, reflecting the presence of VC transitions in this context. We begin by examining word-initial sequences, in the context /#_t/. Having moved on to a different language, we make a few adjustments to the constraint weights in order to keep costs relatively low in this illustration: parameters are w_{σ} = 1,000; w_{seg} = 1,000; w_{MaCo} = 100,000; w_{MiDi} = 500; stop bursts emerge at duration 8; open transitions emerge at duration 11.

(18) Word-initial clusters

Inv. A	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
te	Neutr.	10.75	19.75	17943	0	0	5	0	35886
Inv. B									
te		10.75	19.75	17943					
k˺te	7.5	9.5	18.5	106806	1	-100000	1	8000	32749
Inv. C									
te		10.75	19.75	17943					
kte	8	6.5	18.25	107556	1	-100000	2	4500	29999
Inv. D									
te		10.75	19.75	17943					
k^{h}te	11	5.75	18.25	126265	1	-100000	3	2000	46208

Inventory C, with contrast between /te/ and audibly released /kte/, emerges as optimal given these parameters. This pattern corresponds to Tsou: as noted above, Wright finds that all C1 stops are released in word-initial stop-stop clusters. Consider now what happens word-medially, in the context /V_t/.

(19) Word-internal clusters

Inv. A	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
ate		10.75	19.75	17943					
a<k>te	Neutr.	10.75	19.75	17943	0	0	5	0	35886
Inv. B									
ate		10.75	19.75	17943					
ak˺te	7.5	9.5	18.5	106806	1	-100000	2	1500	26249
Inv. C									
ate		10.75	19.75	17943					
akte	8	6.5	18.25	107556	1	-100000	3	1000	26499
Inv. D									
ate		10.75	19.75	17943					
ak^{h}te	11	5.75	18.25	126265	1	-100000	4	500	44708

In this context, /ak˺te/ with unreleased C1 emerges as optimal. With the extra cues available from the preceding vowel, the effects of compression emerge, favoring the less durationally-marked candidate inventory.

As it stands now, the grammar is categorical in its predictions for C1 realization: audibly released word-initially, unreleased word-internally. In reality, the word-internal realization is probabilistic: about 40% of C1 stops here are released. We could build probabilistic phenomena into the model by changing our simplified assumptions about how release is related to duration. In the current model, an audible burst emerges 'automatically' at some fixed duration threshold for C1. A more realistic model would reflect the fact that a burst becomes incrementally more salient as the segment is lengthened. These probabilities could be reflected in the evaluation of MiDi. For instance, if word-internal C1 with duration of 7.5 units has a 40% chance of being audibly released (perceptual distance 3) and a 60% chance of being unreleased (distance 2), the

perceptual distance between that stop and its absence could be construed as a weighted average, distance 2.4. This would allow the grammar to settle on candidate inventories that have a reasonable chance of realizing distinct contrasts, without predicting completely categorical and invariant realizations within each context.

5.5 Some further predictions

All of the phenomena examined so far serve as illustrations of how a licensing-by-cue approach to phonology can be combined with a gradient theory of timing to give a unified account of certain phonetic and phonological phenomena. Given a minimally adequate theory of timing, these phenomena follow more or less straightforwardly from a phonological approach grounded in phonetics. In such a framework, for instance, a single gradient constraint calling for adjacent segments to overlap would suffice to derive all of the patterns discussed so far (but would not be able to capture incremental compression effects of the sort discussed in the preceding chapters).

Because the formalism here incorporates a very specific theory of timing, it generates additional phonotactic predictions. These mainly involve the possibility of long distance interactions between segments, mediated by higher-level duration constraints. In this section, we examine a set of such predictions, and draw some preliminary conclusions about their empirical plausibility.

5.5.1 Rhotics and clusters in Spanish

Much of the preceding discussion was concerned with repairing certain sequences by increasing the duration of one of the segments (or part of a segment) in that sequence. The cost of doing this comes in the form durational markedness assessed by higher-level duration targets. In most of the asymmetric cases, the difference was between one context where there is pressure to lengthen a stop in order to achieve open transition, and one context where there is less pressure.

Even across contexts that include equal pressure to lengthen a stop, however, we predict that there could be differences in a stop's *ability* to lengthen sufficiently. This would depend on factors such as the duration of other segments in the sequence or the tendency of those segments to induce compression. In this section, I argue that certain phenomena in Spanish instantiate this prediction. In contexts where realizing an open transition will help with stop perceptibility, the availability of this repair is governed by properties of the adjacent segment and compression within the syllable.

As a starting point, consider again the analysis of why English doesn't license stop-stop clusters in word-onset position. That table is shown below.

(20) Stop-Stop neutralization in English

Inv. A	*/k/ dur.*	*/t/ dur.*	*/e/ dur.*	*MarkCost*	*Contrasts*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
te		9.5	18.5	27852.5					
k˺te	5.25	5.25	18.5	239265.6	1	-91500	3	200	175818
Inv. B									
te	Neutr.	9.5	18.5	27852.5	0	0	--	0	55705
Inv. C									
te		9.5	18.5	27853					
kte	8	5.25	18.5	1510453	1	-91500	2	300	1447106
Inv. D									
te		9.5	18.5	27853					
k^{h}te	11	5.25	18.5	4639453	1	-91500	3	200	4576006

Inventory D, with an open transition between the two segments in the cluster, is sub-optimal here because it incurs high durational markedness costs. The long duration of C1 results in a severe violation of the syllable duration constraint. This could be remedied if it were possible to shorten the other segments enough to ameliorate the violation of the syllable target, but the CCV item in Inventory D has already compressed C2 and the vowel close to their respective floor durations.

If C2 were able to crowd further into the vowel, however, it might be possible for a candidate like Inventory D to win. For this to happen, we would need to shorten C2, crowd it into the vowel, and lengthen C1 to create an open transition, all without violating the syllable-duration constraint too severely. For C2 to crowd further into the vowel, it would need to either have a shorter inherent duration or a very high vowel-recoverability coefficient *k*.

While English doesn't have very short stops that carry a lot of vowel information, Spanish has a segment that fits this description to some extent: the rhotic tap /ɾ/. This segment is very short, and in clusters it is realized with a special feature that carries a lot of information about an adjacent vowel. This feature, to be discussed below, is one difference between the Spanish and English versions of the tap.

Ladefoged (1971) defines a ballistic tap as "formed by a single contraction of the muscles such that one articulator is thrown against the other." In apical taps, the tip of the tongue is thrown against the alveolar ridge, producing a very short period of closure, followed almost immediately by release. American English post-tonic intervocalic coronal stops (e.g. *butter*) are ballistic taps. Given the acoustic nature of this segment, which is an extremely brief cessation of energy in the speech signal (essentially a very short stop), surrounded by formant transitions, it would not seem to be a good context in which to realize a stop. Stops, recall, are best perceived adjacent to segments with high sonority.

Nevertheless, stop-/ɾ/ clusters occur in Spanish. They are repaired by realizing an open transition between the two segments, in the form of an excrescent vowel. This is illustrated below with an utterance from a Venezuelan native speaker.

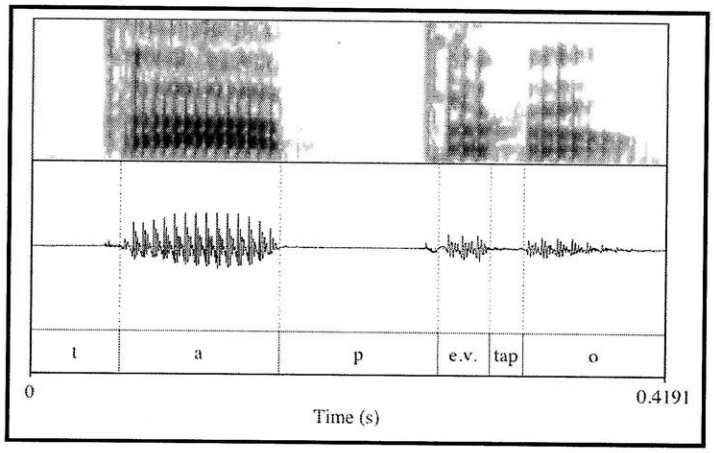

Figure 5.5. *Utterance of nonce-word /tapro/ from a Venezuelan Spanish speaker. Clearly visible in the waveform and spectrogram is an excrescent vowel (labeled 'e.v.') separating [p] and [ɾ].*

There are several properties of this sequence that help maximize cues to the presence and features of all three segments. First, the presence of an excrescent vowel in between the two consonantal occlusions guarantees that C1 will have a clear burst that is not masked by the tap, while also providing formant transitions that could be cues to the presence and features of both consonants. Second, the excrescent vowel partially tracks the formants of the following vowel, so it should aid recoverability of that vowel. In fact, Bradley & Schmeiser (2003), building on earlier work by Gili Gaya (1921), Steriade (1990), Hall (2003), and Bradley, propose that the excrescent vowel *is* simply a part of the following vowel: the tap itself has been crowded far enough into that vowel that part of the vowel is audible before the tap constriction is in place.

Acoustic data from two Venezuelan speakers lend some support to this hypothesis. Compared to a singleton tap, a stop-tap cluster does result in incremental shortening of the following vowel. This is shown in figure 5.6. This is in contrast to word-medial stop-stop clusters, which generally

do not contain open transitions and induce no incremental shortening. So there does seem to be some crowding of the tap into the following vowel in a way not associated with other clusters.[9]

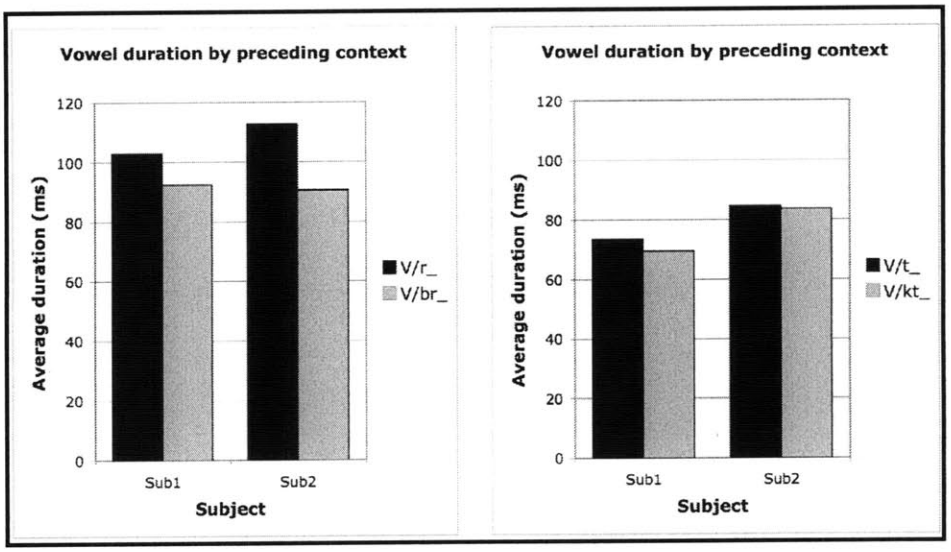

Figure 5.6. *Vowel duration in ms of /o/ for two Venezuelan Spanish speakers in the contexts /tar_/, /tabr_/ (left graph), /tat_/, and /takt_/ (right graph). The stop-tap clusters result in incremental vowel shortening, while the stop-stop clusters result in little or none.*

There is probably more to the story than a simple shift in the timing of the tap, however. Recasens & Pallarès (1999) report that, for rhotic /ɾ/ in the closely-related language Catalan, the area of the tongue just anterior to the dorsum is lowered to create a concave tongue shape. So this sound may be more complicated than a simple ballistic tap. Despite the presence of this tongue-body gesture, the excrescent vowel before a tap does convey information about the

[9] That said, clusters other than obstruent-liquid are arguably heterosyllabic. There may be less pressure to shorten a vowel following /takto/ because it is syllabified as /tak.to/. The motivation for this syllabification presumably would be that /kt/ is not a possible word onset, so it shouldn't be a possible syllable onset either. However, it is equally well the case that /k/ is not a possible word ending, so no independent facts about the language call for such an analysis. In any case, this is the best comparison that the language offers us for our current purposes.

following vowel. Shown below is the formant space for the excrescent vowels of subject 1 from the charts above. This formant space is superimposed on one from stressed vowels and one from the open phases of apical trill /r/ for comparison. The formant space is derived by measuring the average formant values internal to the relevant item in the context of various succeeding vowels. Each point in the space shows the average F1 and F2 values of the relevant item in the context of a particular following vowel.

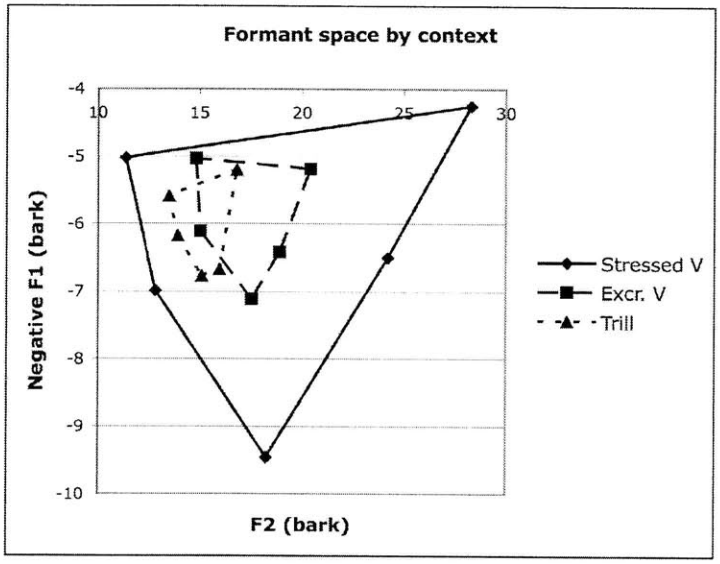

Figure 5.7. *Formant space for one Venezuelan speaker internal to stressed vowels (largest space), excrescent vowels in stop-tap sequences (intermediate space), and the open phases of trilled /r/ (smallest space). Data points within each space are respective average values for {u, i, e, a, o}, clockwise from top left. Y-axis is inverted to preserve height orientation of vowel chart.*

Although the formant space for the excrescent vowel is compressed somewhat relative to a full stressed vowel, it still obviously retains the general shape of the full vowel space. This means that it contains acoustic information that can help the listener to discriminate following vowel contrasts.

Before we can analyze the licensing of stop-tap clusters, there is one additional preliminary to be addressed. In word-initial position, these clusters contrast not with a singleton tap, but with a multiple-cycle apical trill, [r]. For the purposes of this analysis, we assume that [ɾ] results from shortening [r] past some threshold. In this view, [ɾ] can be though of as a single-cycle trill or failed trill. [r] is a relatively complex, difficult sound to make; only a narrow range of aerodynamic conditions in the back cavity can sustain tongue-tip vibration (Solé 1999, 2002). The constraints that vibration places on tongue position are also extensive: the posterior dorsum is retracted into the pharynx and the anterior dorsum is lowered to create a concave tongue shape. These movements allow the tongue tip more room for the vertical oscillation inherent to trilling (see Recasens & Pallarès 1999 for Catalan data and further references).

Given the difficulty of producing a sustained apical trill, it seems plausible that creating the necessary aerodynamic conditions requires a certain amount of time. If the trill is shortened too much, it may be physiologically implausible or impossible to create these conditions. In this case, the result is [ɾ]. This sound is acoustically close to a single cycle of the trill: occlusion followed immediately by release. From an articulatory perspective, it can be thought of as a shorter version of the tongue-tip constriction and at least the anterior dorsal gesture associated with [r]. The hypothesis that [ɾ] is a single-cycle or failed trill also offers an account of why [r] and [ɾ] display positional neutralization in Spanish. In a singleton onset, where syllable duration effects are not as pressing, the long form [r] emerges. In clusters, where syllable duration effects are relatively pressing, the short form emerges in a way that helps preserve cues to contrasts in adjacent positions.

Another approach could invoke articulatory markedness and prosodic strengthening to explain these facts. We might say that [r] is dispreferred following stops because it is just harder to produce than [ɾ]. In syllable-initial position, however, it may be desirable to use a stronger or more acoustically disruptive realization of /r/ to help signal a prosodic boundary. This would be in the spirit of Bakovic's (1994) initial strengthening analysis. In fact, the two hypotheses are not mutually exclusive. Any theory probably needs to say something about initial strengthening in order to explain why [r] and [ɾ] contrast between tautomorphemic vowels but neutralize to one or the other realization in other contexts. This would also explain why the contrast neutralizes to something more like a full trill after unambiguously heterosyllabic consonants such as /n/ and /s/; although there is an enormous amount of variation in these contexts, as discussed by Hammond (1999), Lewis (2004), and Bradley (2006), full trill is at least a possible realization.

Figure 5.7 shows that multiple-cycle [r] contains less acoustic information about the following vowel than the excrescent vowel associated with [ɾ]. We represent this distinction as a difference in the vowel-recoverability coefficient *k*. Furthermore, we will assume that [r] 'automatically' becomes a single-cycle trill (notated as [ɾ]) when it shortens past a certain degree. The excrescent vowel will be treated as part of [ɾ] in the sense that it results in a high vowel-recoverability coefficient for [ɾ]. It will also be associated with a preceding stop, because we assume that the excrescent vowel is only audible when that stop is realized with sufficient temporal distance from the tap closure. This will affect the perceptual distance of the singleton-cluster contrast.

Given these assumptions, we can explain why /bɾe/ is a well-formed syllable in Spanish, while */bde/ is not. We start by analyzing stop-stop clusters; the situation here is very much like the analysis of English in (19). Parameters are: $w_{\sigma} = 1{,}000$; $w_{seg} = 1{,}000$; $w_{MaCo} = 81{,}000$; $w_{MiDi} = 4{,}000$; $t_{\sigma} = 30$; $t_{Cons} = 13$; $t_{Vowel} = 25$; $d_{trans} = 4$; $j = 0.4$. We assume that open transitions following a stop emerge at duration 11 and [r] becomes [ɾ] when shortened past 8. The vowel-recoverability coefficient k is set to 0.3 for stops and [r], 0.6 for [ɾ].

(21) Stop-stop neutralization in Spanish

Inv. A	*/b/ dur.*	*/d/ dur.*	*/e/ dur.*	*MarkCost*	*Contr.*	C_{MaCo}	*Dist.*	C_{MiDi}	*InvCost*
de	Neutr.	11.25	17.5	17001	0	0	--	0	17001
Inv. B									
de		11.25	17.5	17001					
bəde	7	8.75	16.25	110538	1	-81000	1	64000	110539
Inv. C									
de		11.25	17.5	17001					
bde	8	8	16.5	112500	1	-81000	2	36000	84501
Inv. D									
de		11.25	17.5	17001					
bəde	11	7	16.75	137265	1	-81000	3	16000	89266

The table above compares four candidate inventories for the context /#_de/. Note that superscript 'ə' indicates an excrescent vowel. Given these weightings, neutralization to /de/ emerges as optimal. Lengthening C1 sufficiently to create good cues to the contrast results in high duration markedness costs, just like our analysis of English above.

To consider the situation with /r/, where the duration of C2 has strong effects on its phonetic realization, we need to expand our inventories to at least three members. This is because producing /r/ as a single-cycle trill presumably could make its presence or absence less salient than a trill realization. To formally reflect this, we need to consider at least the /be/-/bre/ contrast. We make the minimal necessary assumption, that the /be/-/bre/ contrast is more distinct with trill realizations than with tap realizations; we assign a distance of 2 to the tap realizations and 3 to the trill ones. Given the parameters used above, the inventory costs here come out as negative numbers; because the zero-cost point is arbitrary in this formalism, the fact that the numbers are negative has no particular meaning. For simplicity, we don't consider candidates that neutralize /be/ and /re/.

(22) Stop-tap clusters in Spanish

Inv. A	*C1 Dur.*	*C2 Dur.*	*V Dur.*	*Mark Cost*	*Contr.*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
be		11.25	17.5	17001							
rre	Neutr.	11.25	17.5	17001	1	-81000	4	be-re	4000	4000	-42998
Inv. B											
be		11.25	17.5	17001			4	be-re	4000		
rre		11.25	17.5	17001			2	be-bre	36000		
bre	8.5	7.9	14.75	87331	3	-243000	2	re-bre	36000	76000	-45667
Inv. C											
be		11.25	17.5	17001			4	be-re	4000		
rre		11.25	17.5	17001			2	be-bre	36000		
bᵊre	11	7.9	14.25	100581	3	-243000	3	re-bre	16000	56000	-52417
Inv. D											
be		11.25	17.5	17001			4	be-re	4000		
rre		11.25	17.5	17001			3	be-bre	16000		
brre	7	8.75	16.25	110538	3	-243000	2	re-bre	36000	56000	-42460
Inv. E											
be		11.25	17.5	17001			4	be-re	4000		
rre		11.25	17.5	17001			3	be-bre	16000		
bᵊre	11	8	16.5	139500	3	-243000	3	re-bre	16000	36000	-33498

With the same constraint weights as the table above, we derive a contrast between /re/ and a stop-tap cluster with excrescent vowel, Inventory C. The increased vowel-recoverability coefficient of the innermost consonant allows it to crowd into the vowel more, which in turn allows the stop to achieve enough temporal separation from the following tap to maximize cue availability.

In the current formalism, where phonotactic licensing is affected by general principles of timing, we might expect a connection between the licensing of obstruent-/r/ clusters and the licensing of other clusters. To derive the grammar of Spanish, for instance, we posited a relative weighting for MaCo and durational markedness that makes it rather costly to preserve contrasts, such as /de/-/bde/, that add substantial duration to a syllable. Obstruent-/r/ clusters are somewhat exceptional, in the sense that they offer the possibility of preserving cues to contrast while still compressing segments a fair bit.

This general weighting configuration has consequences for other clusters. For instance, consider the situation for /s/-stop clusters. We retain the weights from above, and assume that the distinctiveness of the /pe/-/spe/ contrast gets better as the /s/ gets longer.

(23) /s/-stop clusters in Spanish

Inv. A	*C1 Dur.*	*C2 Dur.*	*V Dur.*	*Mark Cost*	*Contr.*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
se		11.25	17.5	17001							
pe		11.25	17.5	17001	1	-81000	se-pe	4	4000	4000	-42998
Inv. B											
se		11.25	17.5	17001			se-pe	4	4000		
pe		11.25	17.5	17001			se-spe	3	16000		
spe	7	8.75	16.25	110538	3	-243000	pe-spe	1	64000	84000	-14460
Inv. C											
se		11.25	17.5	17001			se-pe	4	4000		
pe		11.25	17.5	17001			se-spe	3	16000		
spe	8	8	16.5	112500	3	-243000	pe-spe	2	36000	56000	-40498
Inv. D											
se		11.25	17.5	17001			se-pe	4	4000		
pe		11.25	17.5	17001			se-spe	3	16000		
spe	11	7	16.75	137265	3	-243000	pe-spe	3	16000	36000	-35733

Given these values, neutralization comes out optimal. This is indeed the grammar of Spanish: there are no word-initial /s/-stop clusters, and all such clusters that entered the language from Latin are repaired by vowel epenthesis, e.g. Latin /studiu/ 'study' becomes Spanish /estudio/ (Lief 2006). Portuguese patterns much the same way.

The analysis as currently stated does not predict that reduction to tap in rhotic clusters always implies non-availability of sC clusters. It may, however, make a related prediction: absence of /s/-stop clusters should always imply absence of rhotic clusters with full trills. The contrapositive, that presence of full trills in rhotic clusters implies licensing of /s/-stop clusters, would also be predicted. This prediction holds only with certain assumptions in place.

First, we assume that clusters with full trills are always at least as durationally marked as /s/-stop clusters. Because full trills are difficult to produce, subject to aerodynamic interference from adjacent segments, and rather long inherently, this assumption seems plausible. If /s/-stop clusters turn out to be much longer than stop-trill clusters in some languages, the prediction would not go through.

Second, we assume that the contrast between singleton stop and /s/-stop is (or can be) at least as distinct as the contrast between trill and stop-trill. More generally, all of the typological predictions made by the current approach rely on the possibility of comparing the distinctiveness of different contrasts. This is fairly straightforward in cases where the cues to two contrasts are in a subset relation, as in the case of released and unreleased stops. It is more difficult to compare contrasts that rely on different types of cues; in the current case, the contrast involving stop-trill

clusters relies largely on external cues to the presence and features of the stop, while the contrast involving /s/-stop clusters relies largely on internal cues to the presence and features of the /s/. Under the assumption that the distinctiveness of different contrasts can be compared through a measure such as confusability in noise, it would be possible to test the relative distinctiveness of these contrasts. If it turned out that the contrast involving stop-trill clusters was more distinct, then the prediction made here would not be justified. We return to this issue in section 5.5.2.

Given these assumptions, the current formalism makes it impossible for a rhotic cluster with a full trill to contrast with a singleton while not allowing /s/-stop to contrast with a singleton stop. To see this, consider candidate inventory E from (22) and D from (23). The two inventories are identical with regard to the number of contrasts preserved and the distinctiveness of those contrasts. The inventory with trill clusters is more durationally-marked than the one with /s/-stop clusters. In order for the inventory with trill clusters to emerge as optimal, the benefit conferred by MaCo must be larger than the sum of durational markedness and the cost assessed by MiDi. Because this number is larger for the inventory with trill clusters than the one with /s/-stop clusters, this entails that some inventory with /s/-stop clusters will also emerge as optimal.

For the inventory with trill clusters to beat out neutralization and clusters with reduced trills, we need to increase the weight of both MaCo (to beat neutralization) and MiDi (to beat the inventory with a shorter and less salient /r/). Increasing those weights to 110,000 and 8,000, respectively, will accomplish this.

(24) Stop-trill clusters

Inv. A	*Mark Cost*	*Contr.*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
be	17001							
rre	17001	1	-110000	4	be-re	8000	8000	-67998
Inv. C								
be	17001			4	be-re	8000		
rre	17001			2	be-bre	72000		
b°re	100581	3	-330000	3	re-bre	32000	112000	-83417
Inv. E								
be	17001			4	be-re	8000		
rre	17001			3	be-bre	32000		
b°re	139500	3	-330000	3	re-bre	32000	72000	-84498

These weights also predict that /s/-stop clusters will be licensed. Inventory D, with a relatively long /s/, wins given these particular weights. This is shown below.

(25) /s/-stop clusters

Inv. A	*C1 Dur.*	*C2 Dur.*	*V Dur.*	*Mark Cost*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
se		11.25	17.5	17001						
pe		11.25	17.5	17001	-110000	se-pe	4	8000	8000	-67998
Inv. B										
se		11.25	17.5	17001		se-pe	4	8000		
pe		11.25	17.5	17001		se-spe	3	32000		
spe	7	8.75	16.25	110538	-330000	pe-spe	1	128000	168000	-17460
Inv. C										
se		11.25	17.5	17001		se-pe	4	8000		
pe		11.25	17.5	17001		se-spe	3	32000		
spe	8	8	16.5	112500	-330000	pe-spe	2	72000	112000	-71498
Inv. D										
se		11.25	17.5	17001		se-pe	4	8000		
pe		11.25	17.5	17001		se-spe	3	32000		
spe	11	7	16.75	137265	-330000	pe-spe	3	32000	72000	-86733

This pattern, with full trills in clusters and /s/-stop clusters allowed word-initially, is exemplified by Italian. Shown below is a male Italian speaker from Milan producing the nonce-word /brano/. Clearly visible are three separate occlusion and release cycles internal to the trill.

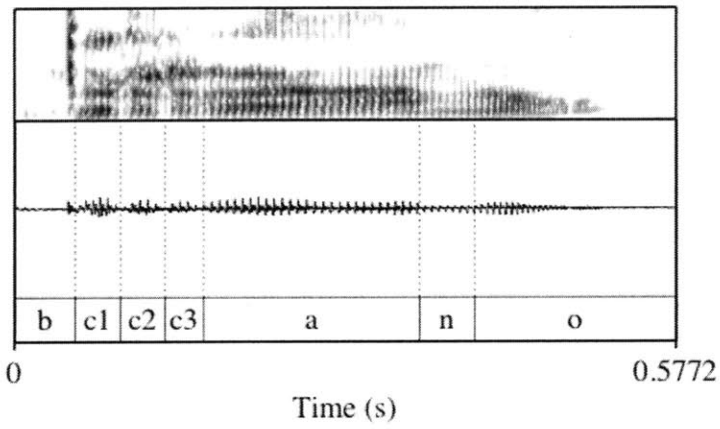

Figure 5.8. *Utterance of nonce-word /brano/ from a native Italian speaker with a multiple vibrant trill. The [r] contains three cycles, labeled c1-3 in the figure.*

This speaker produces the majority of such clusters with full trills. When the trills 'fail', they are produced as fricatives or approximants, occasionally devoiced, but never as taps. It is not generally possible to tell whether preceding stops are separated from following trills by an excrescent vowel, or simply released into the open phase of the trill cycle.

Agreeing with our prediction, Italian allows word-initial /s/-stop clusters, e.g. /spavento/ 'fright', /stesso/ 'same'. Based on an informal survey of Romance languages, Romanian and French seem to pattern with Italian in this regard. Recordings from one female speaker of Romanian show that a full trill is possible in such clusters, although tap may be more common. Accordingly, Romanian allows words such as /stɨŋgʌ/ 'left (hand)'. French features a uvular rather than apical

trill (although not in all varieties); full trill realization in clusters appears to be possible (Haden 1955). French also allows words such as /skandal/ 'scandal' and /spazmə/ 'spasm'.

5.5.2 Onset-coda dependencies I: implicational asymmetries

The general logic of the predictions above concerning /s/-stop and /Cr/ clusters applies to other types of inventories as well. The idea is that, if one type of timing configuration is more durationally marked than another and has no advantage with regard to other factors, then licensing the more marked configuration entails licensing the less marked configuration. This is a strong prediction of the current theoretical approach; it entails that their are hierarchies of contrasts based on their relative durational markedness and perceptual distinctiveness. All languages should conform to these hierarchies: licensing of a more marked contrast entails licensing of a less marked contrast. This general property makes a number of typological predictions; to identify these predictions, however, we need to know which phonetic strings are more temporally marked than others and which phonetic strings result in more distinct contrasts. In practice, we do not usually have this knowledge.

For instance, in languages like English and Spanish, it seems plausible to assume that stop-liquid clusters are less durationally marked than a stop-stop cluster with open transitions would be, and that the cues to singleton-stop clusters are roughly the same in the two strings. It also seems plausible that stop-liquid clusters result in more distinct singleton-cluster contrasts than stop-stop sequences *without* open transitions. Given these assumptions, it follows that licensing of stop-stop clusters entails licensing of stop-liquid clusters. In cases where these assumptions do not

hold, however, such as the pre-stopped liquids with contrastive glottalization in Montana Salish discussed above, the prediction no longer holds. Furthermore, we can't know whether the prediction holds unless we know the phonetic details of liquids (and stops) in a language and the system of contrasts that they participate in.

More generally, it is impossible for the current system to make any predictions about typology without knowledge of the durational patterns at issue and a theory of contrast distinctiveness. In this chapter, we examine several putative typological predictions and attempt to falsify them. All of the predictions themselves are contingent on assumptions about duration and distinctiveness of contrasts. While the theoretical approach developed here in principle makes an almost unlimited number of such predictions, we limit discussion here to a small number of cases that don't require outlandish assumptions about duration and perception.

The /s/-stop and stop-trill asymmetry discussed above holds between two types of clusters in onset position. However, because the timing target is held to be some larger unit such as a syllable, we predict similar asymmetries between phonological strings at different positions in that larger unit. So, for instance, any equally perceptible contrast that relies on less durationally marked configurations than a stop-stop cluster with open transitions should be licensed in languages where stop-stop clusters with open transitions are licensed. One such contrast might be the presence of a coda consonant: because a singleton coda stop has transitions from a preceding vowel, it should need to lengthen its burst less than a preconsonantal stop in order to achieve the same level of perceptual robustness. Because the coda stop can impinge upon a preceding vowel,

it should also result in overall shorter syllables than an initial stop-stop cluster with open transition.

We predict, then, that any language which licenses onset stop-stop clusters with open transitions should also license coda singleton stops.[10] To see this, first consider a set of weights that licenses both types of contrast. For the purpose of this simulation, we treat a singleton coda stop identically to an onset one. All parameter values are carried over from the Romance examples, except for the contrast constraints: w_{MaCo} = 100,000; w_{MiDi} = 4,000.

(26) Complex onsets and codas

Inv. A	*Coda Dur.*	*C1 Dur.*	*C2 Dur.*	*V Dur.*	*Mark Cost*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
de			11.25	17.5	17001		de-deb	3	16000		
deb	8		8	12.5	112500		de-bde	3	16000		
bəde		11	7	16.75	137265	-300000	deb-bde	4	4000	36000	2766
Inv. B											
de			11.25	17.5	17001						
deb	8		8	12.5	112500	-100000	de-deb	3	16000	16000	45501
Inv. C											
de			11.25	17.5	17001						
bəde		11	7	16.75	137265	-100000	de-bde	3	16000	16000	70266

[10] Kaye & Lowenstamm (1981) make a related but stronger prediction: every language that allows branching (complex) onsets should also allow branching rimes (i.e., codas and glides). All of the languages discussed below could conceivably be problematic for this generalization, although I will argue that some of them can be analyzed as containing codas. Chiquihuitla Mazatec, though, would seem to be a clear counterexample. This language allows word-initial sequences such as /rk/ (where /r/ is described as a tap) that are clearly clusters, but appears to have no codas.

Compare now the candidates that collapse the /de/-/bᵊde/ contrast (inventory B) and the /de/-/deb/ contrast (inventory C). They fare equally well on the contrast constraints, but inventory C is more durationally marked. This means that inventory B harmonically bounds inventory C. If we decrease the weight of MaCo to select a neutralized inventory, that neutralized inventory will always retain the coda stop over the onset stop with open transition.

Possible counterexamples to this prediction are languages that license complex onsets with open transitions, but do not license codas. We are aware of seven languages that have been described as licensing CCV but no CVC syllables: Arabela, Cheke Holo, Lakhota, Mazateco, Pirahã, Piro, and Tsou. We briefly examine evidence about syllabification and transition quality in each of these languages.

Arabela licenses only obstruent-glide sequences in onset position, which presumably do not contain open transitions (Rich 1963). Furthermore, glides appear to be licensed in coda position, although these sequences may be analyzed as diphthongs.

Cheke Holo allows only obstruent-liquid clusters (White *et al.* 1988). No phonetic description is available, but again, we at least expect that such clusters do not *require* an open transition to make the stop perceptible. Additionally, Cheke Holo allows word-medial heteroorganic nasal clusters, which are not licensed word-initially; these may be analyzed as heterosyllabic (Blevins 2006).

Lakhota does allow stop-stop clusters word-initially, and they are realized with open transitions (Albright p.c.). Although word-final consonants are not generally permitted, some do surface in function words and in casual speech, including voiceless obstruents (Albright 2004). In fact, Albright proposes that codas are banned only in roots. Whatever is driving this morphological difference, we would need to say that it is not the grammar of timing in order to preserve the generalization at issue here.

Mazateco complex onsets can largely be analyzed as complex segments with secondary articulations. In the Chiquihuitla variety, those that are clearly clusters are /s/-obstruent and /rk/, where /r/ is described as a voiceless tap (Jamieson 1977); neither is described as having an open transition. Even if the tap-/k/ clusters do contain an open transition, it might not be problematic with regard to the prediction above. /r/ is presumably described as a tap in part because it is very short; as such, it wouldn't necessarily be more durationally marked than a consonant in coda position.

Pirahã has no complex onsets on the surface. It was only classified as such because one analysis has surface aspirated stops deriving from geminate onset stops (Blevins 1995).

Piro allows stop-stop onsets word initially. From the description in Matteson (1965), there appear to be open transitions in these clusters. Word-medial clusters are analyzed as complex onsets, but only because there are word-initial clusters and no word-final consonants. The implicit argument would be that treating these medial sequences as complex onsets gives a unified account of the syllable- and word-level phonotactics. However, this is not correct. There

are various word-medial clusters that do not occur word-initially, e.g. /hinmunami/ 'tree species', /pi͡kxka/ 'like, equal to'. Comparison of the cluster chart and dictionary in Matteson (1965) suggests that at least 15 clusters are attested word-medially but not word-initially. This suggests that some word-medial clusters will need to be analyzed as heterosyllabic, with the word-final restriction explained by some other principle. Blevins (1995) notes additionally that vowel length is in a trading relation with the length of a following consonant; Matteson (1965) appears to show that there are cooccurrence restrictions between a vowel and following consonant. Both of these patterns suggest that following consonants are more closely coordinated with a preceding vowel than a following one; if we take the syllable to be a unit of timing, then these facts would support an analysis with codas.

Tsou, as we saw earlier, is ambiguous with regard to word-medial syllabification. In any case, word-initial stop-stop clusters do not normally contain an open transition, as was discussed earlier.

The status of this putative implicational universal, then, is promising if not entirely certain. Piro and especially Lakhota look as if they may present problems, but in each case an analysis of the language as allowing codas is at least possible. We can not solve the problem of why word-final consonants are disallowed in many of these languages, or more general cases where the set of word-final consonants is smaller than that of word-medial codas, as in Spanish. All we can say is that, if these patterns have an explanation grounded in perception or timing, it must pertain to higher-level prosodic factors that have not been investigated here.

5.5.3 Onset-coda dependencies II: cooccurrence restrictions

In addition to predicting that the markedness of equally-perceptible contrasts can differ at different positions in the syllable, the current approach also predicts that the markedness incurred by adding segments to create a contrast is cumulative and increases exponentially across the syllable. In other words, every instance of adding a segment increases the markedness of the syllable, and does so by more than the previous instance. What this means is that we should see effects of doubly-marked structures: cases where two marked structures are independently licensed, but instances of both structures occurring within the same syllable are not.

As a concrete example, consider the licensing of complex onsets and codas. Because the markedness of a CCVC syllable will always be greater than the markedness of a CCV or CVC syllable, we predict that a language could license the latter two types of syllables while prohibiting the former. Such a case is shown below. All paramter values are carried over from the previous examples, except for contrast constraints: $w_{MaCo} = 95{,}000$; $w_{MiDi} = 4{,}000$.

(26) Doubly-marked structure effects

	Coda Dur.	*C1 Dur.*	*C2 Dur.*	*V Dur.*	*Mark Cost*	*Contr.*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
								re-red	3	16000		
Inv. A								re-bre	2	36000		
re			11.25	17.5	17001			re-bred	4	4000		
red	8		8	12.5	112500			red-bre	4	4000		
bre		7	8.75	16.25	110538			red-bred	2	36000		
bred	5.25	5.25	5.25	16.25	300438	6	-570000	bre-bred	3	16000	112000	82477
Inv. B												
re			11.25	17.5	17001			re-red	3	16000		
red	8		8	12.5	112500			re-bre	2	36000		
bre		7	8.75	16.25	110538	3	-285000	red-bre	4	4000	56000	11039
Inv. C												
CV			11.25	17.5	17001							
CCV		7	8.75	16.25	110538	1	-95000	re-bre	2	36000	36000	68539
Inv. D												
CV			11.25	17.5	17001							
CVC	8		8	12.5	112500	1	-95000	re-red	3	16000	16000	50501
Inv. E												
CV			11.25	17.5	17001	0	0				0	17001

Given these weights, inventory B comes out as optimal. This candidate allows complex onsets and codas, but never in the same syllable. There is a small amount of evidence that this may be a well-formed grammar. Such a grammar exists as a stage in the acquisition path of children learning Dutch (Levelt *et al.* 2000). It also appears to be similar to the grammar of colloquial Bamana (Green & Davis 2010). In this language, codas and complex onsets are created by vowel deletion, e.g. /seli/ 'prayer' surfaces as /sel/ and /kɔrɔ/ 'old' surfaces as /krɔ/. In compounds, multiple deletions may occur in cases where they result in CCV.CCV sequences, e.g. /bila/ 'to accompany' + /sira/ 'road' surfaces as /bla.sra/ 'to travel a short distance with someone'. But crucially, multiple deletions are blocked in cases where it would create a CCVC syllable, e.g. /kɔrɔ/ 'old' + /muso/ 'woman' surfaces as /kɔrɔm.so/, */krɔm.so/ 'old woman'. Although this is the only such language we know of, it is not necessarily the case that researchers seeking to describe the syllable structure of some language would be looking for cooccurrence restrictions of this sort.

The effects of cumulative markedness are not limited to this particular case. The theory would make exactly the same prediction about the cooccurrence of complex onsets and complex codas. This also exists as a stage in the Dutch acquisition data (Levelt *et al.* 2000). Albright (2008) finds this as a gradient restriction in English, and Levelt & Van de Vijver (2004) suggest there may be a similar restriction at work in Dutch. Nonetheless, we know of no adult grammar that displays such a restriction as a categorical effect.

The number of doubly-marked structure effects predicted by this theory is enormous. In principle, we could analyze a case where, for instance, six consonants are licensed in the onset of

an open syllable, but only five are licensed in the onset of a closed syllable. Even identifying all of the predictions made in this regard, much less testing them, is quite difficult. The type of prediction exemplified by the five-consonant scenario described above, at least, is not to our knowledge attested in any language.

Finally, there are some long-distance dependencies predicted by the theory that pertain to the inherent duration of one of the segments. For instance, the theory predicts that a language might license complex onsets before short vowels, but ban them before long vowels. This is shown below. Deriving this pattern requires some changes to parameter settings: w_1 (syllable constraint) = 5,000; w_2 (consonant constraint) = 1,000; t_σ (syllable target) = 30; $t_{x,y}$ (consonant target) = 13; d_t (transition duration) = 4; j (vowel recoverability coefficient for transition) = 0.4; vowel floor is 16; consonant floor is 5; w_{MaCo} = 72,000; w_{MiDi} = 4,000. The long vowel /a/ is given a target duration of 28 units, while the short vowel /i/ is given a target of 22.

(27) Another doubly-marked structure effect

	C1 Dur.	*C2 Dur.*	*V Dur.*	*Mark Cost*	*Contr.*	C_{MaCo}	*Pair*	*Dist.*	C_{pair}	C_{MiDi}	*InvCost*
							di-da	3	16000		
Inv. A							di-bdi	2	36000		
di		11.5	15	7303			di-bda	4	4000		
da		9	18	53490			da-bdi	4	4000		
bdi	6.75	8.5	12	101348			da-bda	2	36000		
bda	5.25	7	15.5	188815	6	-432000	bdi-bda	3	16000	112000	30956
Inv. B											
di		11.5	15	7303			di-da	3	16000		
da		9	18	53490			di-bdi	2	36000		
bdi	6.75	8.5	12	101348	3	-216000	da-bdi	4	4000	56000	2141
Inv. C											
di		11.5	15	7303							
da		9	18	53490	1	-72000	di-da	3	16000	16000	4793
Inv. D											
di		11.5	15	7303	0	0				0	7303

Given these weights, inventory B emerges as optimal. This inventory allows complex onsets before short /i/ but not long /a/. It could be described either as neutralizing vowel distinctions following a complex onset or neutralizing the singleton-cluster distinction before long vowels. It does not correspond to any language, statistical restriction, or acquisition stage that we are aware of. This type of prediction, then, is problematic.

5.5.4 Summary and discussion

In this section, we reviewed several kinds of predictions that emerge from the particular formalization of the timing grammar put forward here and its hypothesized interactions with the phonological grammar. The predictions reviewed here fall broadly into four classes: (1) licensing

of a contrast may depend on asymmetries in the temporal realization of a non-adjacent segment; (2) licensing of temporally marked contrasts within the syllable implies licensing of less temporally marked contrasts within the syllable, all else being equal; (3) licensing of a contrast may depend on the presence/absence of a non-adjacent segment; (4) licensing of a contrast may depend on the inherent duration of an adjacent or non-adjacent segment.

We argued that prediction (1) is instantiated by Spanish complex onsets, where the availability of open transitions for a stop in a stop-C cluster depends in part upon the cost of allowing the second C to impinge on the following vowel. We tentatively argued that prediction (2) is borne out by the typology of Romance clusters involving rhotics and sibilants, and by the typology of syllable structure and open transitions. We found some evidence for the prediction in (3), involving colloquial Bamana, language acquisition stages, and statistical tendencies within languages. Nonetheless, we were unable to identify languages showing the full range of phenomena that a system with long-distance interactions would predict, and that range of phenomena is considerable. Finally, we were unable to instantiate prediction (4) in any way.

Developing the theory presented here to better account for the empirical range of attested phonological phenomena is likely to involve both clarifications of what that empirical range is and tools for constraining the predictions of the formalism. For instance, previous researchers have noted the possibility of restrictions on doubly-marked structures, and generally noted that there is no evidence for such restrictions adult language (Kaye & Lowenstamm 1981, Levelt & Van de Vijver 2004, Albright *et al.* in press). Now that researchers have focused their attention on these cases, however, such evidence is beginning to emerge. Here we noted a statistical

restriction discovered by Albright (2008) and a categorical restriction discovered by Green & Davis (2010); it may be the case that we will uncover more such phenomena now that we are actively looking for them. Accordingly, it may be premature to constrain our grammatical theories in order to rule out such cases.

The final prediction listed above, exemplified here by neutralization of singleton-cluster contrasts in the presence of an inherently long but not an inherently short vowel, is as far as we know completely unattested, as are other phenomena that would fall under the rubric of this prediction. The example above, using vowels with different inherent durations, would carry over largely unchanged into predictions about vowels that differ in contrastive length. As another example, the theory would predict that a language might neutralize singleton-cluster contrasts in a syllable closed by a long consonant, but not in a syllable closed by a short consonant. All such predictions are problematic. For this reason, we should at least consider ways of constraining the formalism, on the assumption that such languages are not going to emerge from future research.

If we conclude that it is desirable to avoid all of the predictions above, we could simply change our assumptions about the grammar to the view noted above in fn. 3: phonological patterns are expressed by arbitrary and/or abstract formalisms, and the phonetic implementation reflects an effort to realize those phonological patterns with desirable perceptual and articulatory properties. This would allow us to describe all of the patterns above and stipulate that the unattested patterns are unattested because the formal phonology component simply doesn't consider these kinds of long-distance dependencies. As we noted above, this is a costly solution. We would lose most of the gains made by the phonetically-based approach to phonology: that approach holds out the

hope of *explaining* patterns of phonological contrast with reference to independent properties of perception and articulation; treating those patterns as essentially arbitrary is clearly a less parsimonious theory.

We might attempt to avoid only the long-distance predictions by minimally modifying the licensing-by-cue approach to include a constraint on the overlap of adjacent segments. This is essentially the approach taken by Gordon (2001), Jun (2002), and Flemming (2008). With this approach, we can explain why open transitions appear to be marked, but we make no predictions about how the availability of open transitions relates to properties of non-adjacent segments in a phonological string. This approach, however, misses certain generalization, both phonetic and phonological. First, the constraint favoring overlap should be motivated by independent facts about language, i.e., even when a contrast is not at stake, sounds tend to overlap. This is true, but it is not the whole story: the incremental compression effects discussed in the preceding chapters show that the full pattern of duration-trading phenomena are not explained by a simple preference for overlap between adjacent segments; they hold within some larger constituent. It would still be necessary to explain why some temporal patterns affect the availability of contrast while others appear not to. This approach would also be unable to explain the predictions listed as (1) and (3) above. If Spanish allows the overlap constraint to be violated for stop-/r/ clusters, it should allow the same thing for stop-stop clusters. Prediction (3) is on shakier empirical ground, so we may decide it is worth excluding such predictions from the theory, although I've argued that this would be premature. The approach of restricting timing effects to adjacent segments runs the risk, then, of throwing out the baby with the bathwater.

Another possibility would be to restrict the kinds of inventories over which contrasts are evaluated. We argued in section 5.4.2 that this will probably be necessary on independent grounds. This solution turns out to be difficult in practice, and will require an intricate theory of inventory containment.

To eliminate long-distance predictions, for instance, we could supplement our theory of inventory evaluation with a principle we refer to as *vertical integration*: if a contrast between Y and ∅ is licensed in context X_Z, then it is also licensed in all environments that properly contain the string X_Z. For instance, if the contrast between a stop's presence and absence is licensed in the context /#_re/, it is also licensed in the contexts /#_red/ and /#_ra/. It would not necessarily be licensed in the contexts /s_re/ and /#_de/. This solution, however, is likely to leave us with no way of analyzing any of the long-distance interactions that arise in phonology, such as vowel harmony and long-distance dissimilation.

This solution will eliminate both the long-distance predictions and the segment-quality predictions mentioned above. If we wanted to eliminate only the segment-quality predictions in (4), we could propose a different version of this principle, which we refer to as *horizontal integration*: if a contrast between Y and ∅ is licensed in context #X_Z#, then it is also licensed in all contexts #V_W#, where V and W bear some specified similarity relation to X and Z, respectively. This principle says nothing about containment; it simply requires that a contrast in one environment is extended to another environment if the *entire* environments are sufficiently similar. This will obviously require an extrinsic theory of similarity; for our current purposes, this could be something like 'have the same manner features'. This would ensure, for instance,

that if the presence between a stop's presence and absence is licensed in the context /#_red/ it is also licensed in the contexts /#_rad/ and /#_reb/, but not necessarily in the contexts /#s_red/ and /#_redz/.

The main disadvantages of both these theories are that they considerably complicate the formalism and are somewhat *ad hoc*. In order to enforce such principles, it must be a formal property of the theory that either the outcome of some contrast evaluations are known prior to the calculations of other contrast evaluations, or that all contrasts in all environments are globally compared as part of the grammar. This means that the theory will require either a theory of derivation or a global evaluation of the entire set of possible words. We argued earlier that the latter possibility is likely to be computationally intractable. The former possibility might in principle work: given the examples discussed here, for instance, we could set up a contrast-evaluation algorithm that starts with a syllable containing very few segments, compares it to syllables that differ from it in some circumscribed way (such as a string-edit distance of one, to be concrete), and then takes the output items of that evaluation as the inputs to the next iteration. At every step, the global principles sketched above would be enforced, with priority over inventory-internal contrast constraints.

Needless to say, this considerably complicates the theory, even when we only consider the schematic description given here. Because the empirical facts are uncertain, we leave both the question of how the formalism should be constrained and the precise formal implementation of the eventual answer for future research. What has been accomplished here is the development of

a theory, the generation of predictions from that theory, and a preliminary attempt to evaluate those predictions.

5.6 Conclusion

In this chapter, we developed a framework for modeling the interaction of the timing grammar with phonotactic licensing. The framework was illustrated with examples from several languages pertaining to the timing of consonant clusters and the licensing of those clusters.

Phenomena involving contrast and neutralization are analyzed by way of constraints on contrast, following Flemming (2001). These constraints favor candidate inventories with more contrasting sounds over those with fewer, and contrasts with greater perceptual distance between the members over those with less. The relative weights of those contrast contraints and duration target constraints determine the extent to which unmarked temporal patterns can be altered to 'repair' perceptually weak contrasts.

This framework helps explain why some repairs are not available in some languages: we argued that stop-stop sequences can not be repaired by temporal separation to create an open transition in English or Spanish because this would incur too much cost from the higher-level duration constraint independently proposed for the analysis of compression effects. Without a theory of duration and temporal markedness, it would be difficult or impossible to explain why there are restrictions on temporal repair strategies.

We reviewed several phenomena discovered by other researchers where the perceptual properties of contrasts directly affect the temporal realization of phonetic forms. These pertain to stop-stop clusters in Georgian (Chitoran 1998 *et seq.*) and Tsou (Wright 1996). In both cases, the temporal separation between stops in sequence shows asymmetries that depend on perceptual properties of the strings in question. Where the context or the articulatory properties of the segments demand a high degree of separation for good cues to emerge, they are realized with a high degree of separation. In contexts where cues can be preserved with less separation, we see compression effects re-emerge: less separation is observed. This falls out naturally from a theory where temporal properties are shaped by perceptual considerations.

In Spanish, we saw a case where the availability of open transitions is affected by the realization of an adjacent segment and the ability of other segments in the string to compress or overlap with each other. This is another situation where the availability of a temporal repair seems to be affected by perceptual properties of the string in which it appears. This particular pattern can be explained by the timing theory proposed here, but not by timing theories that posit only constraints governing the overlap of adjacent segments. We also argued that the facts about rhotic clusters in Spanish bear a logical relation to facts about other types of clusters, using data from other Romance languages to explore those logical relations.

The Spanish phenomena pertain to one type of prediction that is generated by the particular theory of timing proposed here, but not necessarily by other approaches in the licensing-by-cue tradition. We proceeded to examine a range of other predictions with this property. The empirical evidence bearing on these predictions was a mixed bag: some appear to be supported, some

appear to describe rare or under-attested phenomena, and some appear to be completely unsupported. We discussed ways that the theory might be constrained to obtain a better empirical fit with the world's languages, but in the end put off a formal statement of these constraints until the empirical picture is better understood.

The approach taken here was to illustrate a few phenomena in some detail, as a demonstration of how this approach might work more generally. There should in principle be many other cases of duration-sensitive perceptual repairs and perception-sensitive temporal patterning. The instances of doubly-marked structure effects examined here, for instance, are just a tiny fraction of all the possible doubly-marked structure effects that could logically exist; the theory presented here predicts more generally that two durationally marked structures may be independently licensed but banned in combination. Other types of contrast should also affect and be affected by timing; for instance, most of what was said here about the licensing of the presence of stops should also apply to the licensing of place contrasts. Most of the cues discussed here (obstruent transients and noise, formant transitions into and out of an obstruent) are also cues to place contrasts. The more general argument made here is that, wherever patterns of phonological contrast are affected by the timing properties of phonological strings, a complete analysis of those patterns requires a theory of why timing patterns have the properties they do and not some other set of properties.

6 Conclusion

This dissertation investigated the grammar of timing as it is reflected in compression effects, and examined how such a grammar might interact with systems of phonological contrast. Here we summarize the results of the investigations and suggest directions for future research.

We began by investigating compression effects, cases where syllables that contain more segments also contain shorter segments, in some detail. An English production study revealed that, while compression effects obtain in a number of contexts, they are not present in every context, and they vary depending on what types of segments are present in a syllable. The principle asymmetry uncovered in this study is that vowels shorten in monosyllabic words with stop-liquid or liquid-stop clusters, relative to their duration in comparable words with singleton liquids; this pattern does not hold in every context for obstruents or nasals, however.

We developed a theory of timing based on weighted, gradiently violable constraints on the duration of segments and syllables. The constraints come into conflict as the number of segments inside a syllable increases, and the weights of the constraints determine what kinds of compression effects will be observed. We showed that this type of grammar can derive the qualitative patterns discovered for English, if the constraints are stated in terms of auditory rather than articulatory representations and if we make certain assumptions about the perceptual properties of various consonants. The general prediction of the model is that vowels shorten more when there is more perceptual information about them in the surrounding context.

A perception study using forward- and reverse-gated stimuli confirmed that most of the perceptual assumptions that the grammar relies on are correct. There is a correlation between those segmental environments that allow more vowel compression and those segments that contain more information about an adjacent vowel. This is a powerful argument that the phonetic representations relevant to temporal coordination encode perceptual properties of speech events.

Although it is not a logical necessity that the grammar of timing affect systems of phonological contrast, we offered a preliminary investigation of what such interaction would predict about phonological systems. The general approach preserves the insights of previous phonetically-based approaches to phonology, which relied on a more minimal theory of timing that essentially calls for adjacent segments to overlap. It also generates new predictions that stem from the details of the particular timing grammar developed here. Some of these predictions correspond to attested phonological patterns that would be difficult to analyze in previous approaches, but are straightforward given the current theory. This approach, however, also predicts entire categories of long-distance phonological dependencies that appear to be unattested in the world's languages. We outlined several approaches to constraining the current theory, pending clarification of the empirical facts.

Many of the questions raised here will require extensive cross-linguistic research to be answered. The timing grammar developed in chapter 3, in particular, is based on English data; although we summarized previous findings in a variety of languages, none of the studies surveyed were comprehensive enough to give us a full picture of compression effects in other languages. It would clearly be useful to examine compression patterns more closely in a variety of languages,

particularly those languages where vowel-length constrasts interact with the number or manner of adjacent consonants.

Another domain where compression effects might be of particular interest is sonority-based phonotactic licensing. If Wright (2004) is correct that the sonority sequencing principle has its roots in perceptibility concerns, and if the current study is right that compression is also related to perceptibility concerns, then we predict that compression may well differ between strings that obey the sonority sequencing principle and strings that do not. Languages such as Georgian and Russian, which contain a wide variety of both types of phonological strings, might provide valuable evidence on which factors affect compression and how these patterns interact with higher-level units such as syllables.

The exploration of phonological implications in chapter 5 was in some ways inconclusive. It seems clear that a theory in which the flow of information between timing grammar and phonological grammar is unconstrained will overgenerate, but it is not entirely clear how much it overgenerates. Some of the long-distance dependencies predicted by such a grammar are completely unattested as a class; for instance, the prediction that availability of complex onsets may interact with differences in the inherent duration of a vowel or coda consonant. Other long-distance dependencies, however, which might appear equally unlikely to a phonologist, do appear to exist. We argued here that colloquial Bamana (Green & Davis 2010) exemplifies just such a long-distance dependency. One (enormous) task for future research, then, is to clarify what types of long-distance dependencies between licensing of phonological contrasts exist. Although we will never fully accomplish this goal, we can at least hope to learn more.

Learning more about the existent patterns of long-distance dependency, in turn, will allow us to constrain the phonological formalism to a suitable degree. We argued that the approach to phonological contrast explored here accomplishes enough that we don't want to discard it, and we briefly outlined a number of ways that approach might be constrained. Which constraints on the approach result in the best theory will depend on what types of languages are attested. With a better understanding of which patterns exist, we can propose concrete measures to constrain the formal approach developed here.

References

Albright, A. (2004). The Emergence of the Marked: Root-Domain Markedness in Lakhota. Presented at the LSA Annual Meeting, Boston, January 2004.

Albright, A. (2008). A universally gradient co-occurrence restriction? Talk presented at the 16th Manchester Phonology Meeting, May 2008.

Albright, A., G. Magri & J. Michaels. (In press). Modeling Doubly Marked Lags with a Split Additive Model. Proceedings of BUCLD 32.

Alwan, A., S. Narayanan & K. Haker. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *Journal of the Acoustical Society of America* **101**(2), 1078-1089.

Baayen, R., D. Davidson & D. Bates. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59**, 390-412.

Bakovic, E. (1994). Strong onsets and Spanish fortition. In Giordano & Ardron (eds.), *MIT working papers in linguistics* **23**: 21-39.

Bates, D. (2007). lme4: An R package for fitting and analyzing linear, nonlinear and generalized linear mixed models. Software application.

Bates, D. (2008). Fitting Mixed-Effects Models Using the lme4 Package in R. Presented at the International Meeting of the Psychometric Society, June 2008.

Beddor, P. (1993). The perception of nasal vowels. In Huffman & Krakow (eds.), *Nasals, nasalization, and the velum* (171-196). New York: Academic Press.

Blevins, J. (1995). The syllable in Phonological Theory. In Goldsmith (ed.), *The Handbook of Phonological Theory* (206-235). Cambridge, Mass.:Blackwell.

Blevins, J. (2006). Syllable typology. In Brown (ed.), *Elsevier Encyclopedia of Language and Linguistics*, 2nd edition (Vol. 12: 333-337). Oxford: Elsevier.

Boersma, P. & D. Weenink. Praat: doing phonetics by computer. Software program.

Bonneau, A. (2000). Letter to the Editor: Identification of vocalic features from French stop bursts. *Journal of Phonetics* **28**, 495-502.

Bradley, T. (2006). Phonetic Realizations of /sr/ Clusters in Latin American Spanish. In Díaz-Campos (ed.), *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish.* Somerville, MA: Cascadilla Proceedings Project.

Bradley, T. & S. Schmeiser. (2003). On the phonetic reality of Spanish /r/ in complex onsets. In Kempchinsky & Piñeros (eds.), *Theory, Practice, and Acquisition* (1-20). Somerville, Mass.: Cascadilla Press.

Broselow, E. (1976). *The phonology of Egyptian Arabic.* PhD dissertation, University of Massachusetts Amherst.

Broselow, E., S. Chen & M. Huffman. (1997). Syllable weight: convergence of phonology and phonetics. *Phonology* **14**, 47-82.

Browman, C. & L. Goldstein. 1986. Towards an Articulatory Phonology. *Phonology* **3**, 219-252.

Browman, C. & L. Goldstein. 1990. Tiers in Articulatory Phonology, with some Implications for Casual Speech. In Kingston & Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (341-397). Cambridge, UK: Cambridge University Press.

Browman, C. & L. Goldstein. 1992. Articulatory phonology: an overview. *Phonetica* **49**, 155-180.

Browman, C. & L. Goldstein. (1995). Gestural Syllable Position Effects in American English. In Bell-Berti & Raphael (eds.), *Producing Speech: Contemporary Issues* (19-33). New York: AIP Press.

Browman, C. & L. Goldstein. 2000. Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée* **5**, 25-34.

Byrd, D. (1994). *Articulatory Timing in English Consonant Sequences.* UCLA PhD Dissertation.

Chitoran, I. (1998). Georgian harmonic clusters: phonetic cues to phonological representation. *Phonology* **15**, 121-141.

Chitoran, I. (1999). Accounting for sonority violations: the case of Georgian consonant sequencing. *Proceedings of the XIV International Congress of Phonetic Sciences*, San Fransisco, 1999, 101-104.

Chitoran, I., L. Goldstein & D. Byrd. (2002). Gestural overlap and recoverability: Articulatory evidence from Georgian. In Gussenhoven & Warner (Eds.), *Laboratory Phonology 7* (419-448). Berlin: Mouton de Gruyter.

Clements, G. & S. Hertz. (1996). An integrated model of phonetic representation in grammar. In Lavoie and Ham (eds.), *Working Papers of the Cornell Phonetics Laboratory No. 11* (43-116). Ithaca, NY: CLC Publications.

Clements, G. & S. Keyser. (1983). *CV phonology : a generative theory of the syllable.* Cambridge, Mass. : MIT Press.

Crystal, T. & A. House. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America* **72**(3), 705-716.

Crystal, T. & A. House. (1988). Segmental duations in connected-speech signals: Current results. *Journal of the Acoustical Society of America* **83**(4), 1553-1573.

Crystal, T. & A. House. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America* **88**(1), 101-112.

Flemming, E. (1995). *Auditory representations in phonology.* PhD dissertation, UCLA.

Flemming, E. 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18: 7-44.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In Hayes, Kirchner, & Steriade (eds.), *Phonetically-Based Phonology* (232-276). Cambridge, UK: Cambridge University Press.

Flemming, E. (2008). The Realized Input. Ms., Massachusetts Institute of Technology.

Flemming, E., P. Ladefoged, & S. Thomason. (2007). Phonetic structures of Montana Salish. *Journal of Phonetics* **36**, 465-491.

Fowler, C. (1981). A Relationship between Coarticulation and Compensatory Shortening. *Phonetica* **38**, 35-50.

Fowler, C. (1983). Converging Sources of Evidence on Spoken and Perceived Rhythms of Speech: Cyclic Production of Vowels in Monosyllabic Stress Feet. *Journal of Experimental Psychology* **112**(3), 386-412.

Fujimura, O. (1987). A Linear Model of Speech Timing. In Channon & Shockey (Eds.), *In Honor of Ilse Lehiste* (109-124). Dordrecht: Foris Publications.

Gili Gaya, S. 1921. La r simple en la pronunciación española. *Revista de Filología Española* **8**, 271-280.

Goldstein, L., D. Byrd & E. Saltzman. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In Arbib (Ed.), *From Action to Language: The Mirror Neuron System* (215-249). Cambridge, UK: Cambridge University Press.

Gordon, M. (2001). Laryngeal timing and correspondence in Hupa. In *UCLA Working Papers in Phonology* **5**, 1-70.

Green, C. & S. Davis. (2010). Avoiding multiple complexities in the prosodic word: Minimization in Colloquial Bamana. Talk presented at the 18th Manchester Phonology Meeting, May 2010.

Guenther, F., C. Espy-Wilson, S. Boyce, M. Matthias, M. Zandipour & J. Perkell. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America* **105**(5), 2854-2865.

Haden, E. (1955). The uvular r in French. *Language* **31**(4), 504-510.

Hall, N. (2003). *Gestures and segments: vowel intrusion as overlap.* PhD dissertation, University of Massachusetts Amherst.

Hammond, R. 1999. On the non-occurrence of the phone /˜r/ in the Spanish sound system. In Gutiérrez-Rexach & Martínez-Gil (eds.), *Advances in Hispanic Linguistics* (135-151). Somerville, Mass.: Cascadilla Press.

Hayes, B., R. Kirchner & D. Steriade (eds.). (2004). *Phonetically Based Phonology.* Cambridge, UK: Cambridge University Press.

Hertz, S. (1991). Streams, Phones, and Transitions: Toward a New Phonological and Phonetic Model of Formant Timing. *Journal of Phonetics* **19**, 91-109.

Honorof, D., & C. Browman. (1995). The center or edge: how are consonant clusters organised with respect to the vowel? In Elenius and Branderud (eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (552-555). Stockholm: KTH and Stockholm University.

Jamieson, A. (1977). Chiquihuitlan Mazatec phonology. In Davis & Poulter (eds.), *Studies in Otomanguean Phonology* (93-106). SIL Publications in Linguistics.

Jun, J. (1995). *Perceptual and articulatory factors in place assimilation: an Optimality-Theoretic approach.* PhD dissertation, UCLA.

Jun, J. (2002). Positional faithfulness, sympathy, and inferred input. Ms., Yeungnam University, Korea.

Katz, J. (2008). English compensatory shortening and phonetic representations. Presented at *Consonant Clusters and Structural Complexity*, Institute of Phonetics and Speech Processing, Munich, July-August 2008.

Kaye, J. & J. Lowenstamm. (1981). Syllable Structure and Markedness Theory. In Belletti, Brandi & Rizzi (eds.), *Theory of Markedness in Generative Grammar* (287-315). Pisa: Scuola Normale Superiore.

Kirchner, R. (1998). *An effort-based approach to consonant lenition.* PhD dissertation, UCLA.

Klatt, D. (1979). Synthesis by Rule of Segmental Durations in English Sentences. In Lindblom & Öhman (Eds.), *Frontiers of Speech Communication Research* (287-300). New York: Academic Press.

Krakow, R. (1999). Physiological organization of syllables: a review. *Journal of Phonetics* **27**, 23-54.

Krull, D. (1990). Relating acoustic properties to perceptual responses: A study of Swedish voiced stops. *Journal of the Acoustical Society of America* **88**(6), 2557-2570.

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics.* Chicago: University of Chicago Press.

Legendre, G., Y. Miyata & P. Smolensky. (1990). Harmonic Grammar – a formal multi-level connectionist theory of linguistic wellformedness: an application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (884–891). Cambridge, Mass.: Lawrence Erlbaum.

Levelt, C., N. Schiller & W. Levelt. (2000). The acquisition of syllable types. *Language Acquisition* **8**, 237–264.

Levelt, C. & R. van de Vijver. (2004). Syllable types in cross-linguistic and developmental grammars. In Kager & Pater (eds.), *Constraints in Phonological Acquisition* (204-218). Cambridge, UK: Cambridge University Press.

Lewis, A. (2004). Coarticulatory effects on Spanish trill production. In Agwuele, Warren, & Park (eds.), *Proceedings of the 2003 Texas Linguistics Society Conference* (116-127). Somerville, Mass.: Cascadilla Proceedings Project.

Lief, E. (2006). *Syncope in Spanish and Portuguese: The diachrony of Hispano-Romance phonotactics*. Cornell PhD dissertation.

Lindblom, B. & K. Rapp. (1973). *Some Temporal Regularities of Spoken Swedish.* Publication 21 of Papers in Linguistics from the University of Stockholm (PILUS).

Luce, R. (1963). Detection and recognition. In Luce, Bush, & Galanter (eds.), *Handbook of Mathematical Psychology* (vol. 1: 103-189). New York: Wiley & Sons.

Maddieson, I. (1985). Phonetic cues to syllabification. In Fromkin (ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (203-221). NY: Academic Press.

Matteson, E. (1965). *The Piro (Arawakan) Language.* University of California Publications in Linguistics 42. Berkeley: University of California Press.

Max, L. & P. Onghena. (1999). Some issues in the statistical analysis of completely random and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research* **42**, 261-270.

Munhall, K., C. Fowler, S. Hawkins, & E. Saltzman. 1992. "Compensatory Shortening" in monosyllables of spoken English. *Journal of Phonetics* **20**, 225-239.

Myers, S. (1987). Vowel shortening in English. *Natural Language and Linguistic Theory* **5**, 485-518.

Nam, H., L. Goldstein & E. Saltzman. (2009). Self-organization of syllable structure. In Pellegrino, Marisco, & Chitoran (eds.), *Approaches to phonological complexity* (297-328). Berlin: Walter de Gruyter.

Newman, S. (1944). *The Yokuts Language of California*. Viking Fund Publications in Anthropology, no. 2. New York: Viking Fund.

Parker, E. & R. Diehl. (1984). Identifying vowels in CVC syllables: Effects of inserting silence and noise. *Perception & Psychophysics* **36**(4), 369-380.

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

Prince, A., and P. Smolensky. (1993). Optimality Theory: Constraint interaction in generative grammar. Ms., Rutgers University, University of Colorado.

Proctor, M. (2009). *Gestural Characterization of a Phonological Class: the Liquids*. PhD Dissertation, Yale University.

Pukui, M. & S. Elbert. (1971). *Hawaiian Dictionary*. Honolulu: University of Hawaii Press.

Quené, H. & H. van den Bergh. (2004). On Multi-Level Modeling of data from repeated measures designs: A tutorial. *Speech Communication*, **43**(1-2), 103-121.

Quené, H. & R. Port. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* **62**(1), 1-13.

Raaijmakers, J., J. Schrijnemakers & F. Gremmen. (1999). How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misperceptions and Alternative Solutions. *Journal of Memory & Language* **41**(3), 416-426.

Recasens, D. & M. Pallarès. (1999). A study of /r/ and /r/ in the light of the "DAC" coarticulation model. *Journal of Phonetics* **27**, 143-169.

Repp, B. & H. Lin. (1989). Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America* **85**(1), 379-396.

Rich, F. (1963). Arabela phonemes and high-level phonology. In Elson (ed.), *Studies in Peruvian Indian Languages: I* (193-206). Norman, OK: SIL Publications in Linguistics and Related Fields.

Rogers, C. & A. Lopez. (2008). Perception of silent-center syllables by native and non-native English speakers. *Journal of the Acoustical Society of America* **124**(2), 1278-1293.

Silverman, D. (1995). *Phasing and Recoverability*. UCLA PhD Dissertation.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, J. Pierrehumbert, J. Hirschberg & P. Price. (1992). TOBI: A Standard Scheme for Labeling Prosody. *Proceedings of the International Conference on Spoken Language 92*, Banff, October 1992.

Smits, R., N. Warner, J. McQueen & A. Cutler. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America* **113**(1), 563-574.

Solé, M. (1999). Phonological universals: Trilling, voicing, and frication. *Berkeley Linguistics Society* **24**, 427-442.

Solé, M. (2002). Aerodynamic characteristics of trills and phonological patterning. *Journal of Phonetics* **30**, 655-688.

Sproat, R. & O. Fujimura. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* **21**, 291-311.

Steriade, D. (1990). Gestures and autosegments: Comments on Browman and Goldstein's paper. In Beckman & Kingston (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (382-397). Cambridge, UK: Cambridge University Press.

Steriade, D. (1997). Phonetics in phonology: the case of laryngeal neutralization. Ms., UCLA.

van Santen, J. (1992). Contextual effects on vowel duration. *Speech Communication* **11**, 513-546.

Waals, J. (1999). *An Experimental View of the Dutch Syllable*. Utrecht University PhD Dissertation.

Warner, N., R. Smits, J. McQueen, & A. Cutler. (2005). Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception. *Speech Communication*, **46**(1), 53-72.

Whalen, D. (1983). Vowel information in postvocalic fricative noises. *Language and Speech* **26**(1), 91-100.

White, G., F. Kokhonigita & H. Pulomana. (1988). *Cheke Holo (Maringe/Hograno) Dictionary*. Canberra: Pacific Linguistics.

Winitz, H., M. Scheib, & J. Reeds. (1972). Identification of Stops and Vowels for the Burst Portion of /p, t, k/ Isolated from Conversational Speech. *Journal of the Acoustical Society of America* **51**(4), 1309-1317.

Wright, J. (1986). The behavior of nasalized vowels in the perceptual vowel space. In Ohala & Jaeger (eds.), *Experimental phonology* (45-67). Orlando: Academic Press.

Wright, R. (1996). *Consonant clusters and cue preservation in Tsou.* PhD dissertation, UCLA.

Wright, R. (2004). A review of perceptual cues and cue robustness. In Hayes, Kirchner, & Steriade (eds.), *Phonetically Based Phonology* (34-57). Cambridge, UK: Cambridge University Press.

Yeni-Komshian, G. & S. Soli. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *Journal of the Acoustical Society of America* **70**(4), 966-975.

Zhang, J. (2004). The role of language specific durational patterns in contour tone distribution. In Hayes, Kirchner, & Steriade (eds.), *Phonetically-Based Phonology* (157-190). Cambridge, UK: Cambridge University Press.