Compression effects and perceptual asymmetries

Jonah Katz
Department of Linguistics & Philosophy
Massachusetts Institute of Technology
77 Massachusetts Avenue
Bldg. 32-D808
Cambridge, MA 02139 USA

aCurrent address:
Institut Jean Nicod
Pavillon Jardin
Ecole Normale Supérieure
29, rue d'Ulm
75005 Paris
France
katzlinguist@gmail.com

Abstract
This paper reports the results of two English experiments on timing and perception. The finding is that temporal patterns in production correspond to asymmetries in perception. We argue that these phenomena are best described in terms of auditory rather than articulatory representations. The production data concern compression effects, whereby, in syllables with a greater number of segments, each one of the segments is shorter than in syllables with fewer segments. The first experiment demonstrates that the amount of vowel compression found in English monosyllabic words depends in part on which consonants occur adjacent to the vowel in that word: words with liquids generally allow more vowel compression than words with nasals or obstruents. To explain these asymmetries, we posit a recoverability target for vowels that can be partially satisfied by information about the vowel contained in the surrounding context; differences in compression, then, can be explained by differences in how much vowel information is contained in adjacent segments and transitions. A second study shows that the perceptual assumptions required to make this analysis work are indeed found in English; timing asymmetries in production mirror perceptual asymmetries. The results favor a theory of timing that includes auditory representations over one that includes only articulatory representations.

Keywords: vowel duration, compression, compensatory shortening, closed-syllable vowel shortening, c-center
1.0 Introduction
This paper is concerned with the timing of segmental sequences and the factors in speech perception and production that influence that timing. Previous studies have shown that vowels tend to be shorter in syllables with more segments than syllables with fewer segments, whether those segments appear in onset or coda position (Fowler 1983, Munhall et al. 1992, van Santen 1992, Clements & Hertz 1996, Shaiman 2001 for English; Lindblom & Rapp 1973 for Swedish; Waals 1999 for Dutch; Maddieson 1985 for cross-linguistic survey of closed-syllable shortening). We refer to these phenomena collectively as compression effects.

Compression effects are of interest, among other reasons, because they have been explained with reference to both articulatory (Fowler 1981, Nam et al. 2009) and auditory (Maddieson 1985, Myers 1987, Clements & Hertz 1996, Flemming 2001) representations. The two approaches are driven by different functional considerations, and hence make different predictions about where compression effects should be observed. As such, clarifying the situation with regard to compression may allow us to shed some light on the division of labor between articulatory and auditory representations, a foundational question in phonetics.

In this paper, we report the results of an English production study that examines whether and where compression effects are observed. We argue that the results are best interpreted in terms of auditory rather than articulatory considerations. That interpretation requires certain assumptions about the perceptual properties of various consonants and transitions; those assumptions are tested in a follow-up experiment and broadly confirmed. The picture that emerges is that the possibility of vowel compression depends in part upon the auditory properties of adjacent consonants and transitions: more compression is observed in environments where the surrounding context contains more information about the vowel. A theory with auditory representations can characterize these facts straightforwardly; they are more difficult to characterize in a theory that includes only articulatory representations.

The next section describes a production study on compression effects in English. Section 3 describes a follow-up perceptual study that sought to confirm the explanation given for the production patterns in section 2. Section 4 summarizes the findings and suggests directions for future research.

2 An experimental investigation of compression effects in English
2.1 Introduction
This section describes a pseudo-word production study that tests for vowel compression within English monosyllabic words. The finding is that patterns of compression depend on the number, manner, and syllable position of adjacent consonants. We propose that these differences may be explained by asymmetries in the perceptual properties of the relevant sounds.

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 label 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 as *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 (consonantal) 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.

### 2.1.2 Previous studies

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. Fowler (1981) includes a brief review of literature on this subject before 1981.

Simplex coda CS, often referred to as *closed-syllable vowel shortening*, is widely attested cross-linguistically. Investigations of this pattern 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 other compression effects. Maddieson (1985) contains an extensive review of languages where the phenomenon has been attested.

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 simplex CS for segments in what they call ‘the extended nucleus’ of an English syllable, which includes voiced transitions preceding and following the vowel, following glides, and following liquids. Munhall et al. (1992) find incremental coda CS for obstruent-obstruent clusters. Shaiman replicates this result for the English pseudo-words /pæp/, /pæaps/, and /pæpst/.

Although none of these studies examined 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; /t/ 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. Waals (1999) finds that, in onset position in Dutch, consonants in clusters are shortened relative to singleton counterparts. Compression disproportionately affects higher-sonority segments relative to lower-sonority ones. For vowels, she finds incremental coda CS with all consonant types, and possibly for two vs. three coda consonants. This effect is much larger for vowels preceding /t/ 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.

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 one study failed to find any effect at all. 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.

Before we describe the current study, we briefly review two approaches to explaining compression phenomena.

2.1.3 An articulatory approach
One approach to compression treats it as an epiphenomenon of articulatory gestural coordination (Fowler 1981 et seq., Browman & Goldstein 1990 et seq., Nam et al. 2009). These theories include no mechanism for actively controlling 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 masked 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. Essentially, the vowels form a ‘scaffold’ that can support consonantal constrictions. This is shown below.
Figure 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 activation interval 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 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 language. 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 an adjacent vowel.

This general theory makes different predictions about compression when coupled with a more specific theory of gestural alignment. Articulatory Phonology (henceforth AP) is 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. 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 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 overlap the following vowel gesture more as the number of consonant gestures in the cluster increases. Under the assumption that the activation interval of a vowel gesture remains fixed from one utterance to the next, this will result in acoustic shortening, as shown by the arrows in figure 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

---

1 In the task–dynamic model underlying AP, the relationship between articulation and acoustics is not necessarily straightforward. A number of factors affect the mapping from gestural activation to actual duration, including speech rate and prosodic juncture. The only crucial assumption made here is that, all else being equal, more articulatory overlap between a vowel and consonant will result in a shorter acoustic duration for the vowel. This assumption alone can generate predictions from the model.
gesture in coda position, regardless of how many other consonant gestures might follow it. This is shown in figure 3.

![Diagram showing vowel and consonant gestures in different syllable positions.](image)

**Figure 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 consonantal gestures are added to a coda. Adding further consonantal gestures will not affect the acoustic duration of the vowel.

This approach makes specific predictions about acoustic compression: incrementally greater 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 earlier implementations 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’.

The articulatory approach predicts that compression should obtain between syllables with different numbers of onset consonants, but not coda consonants. They predict that compression arises whenever gestures overlap, regardless of the internal features of those gestures. 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 we should examine both the acoustic steady states of segments and the transitions between them.

2.1.4 An auditory approach
An entirely different approach to compression effects emerges from investigations of closed-syllable vowel shortening (henceforth CSVS). In this phenomenon, vowels in closed syllables are observed to be shorter than vowels in open syllables. CSVS, then, is a specific kind 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, favors long segments for perceptibility; it also enforces a duration target on larger units, to foster rapid and efficient communication and to create at least a tendency towards isochrony. More isochronous syllables may 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. 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. Myers (1987) invokes this trading approach in a phonological analysis of English closed-syllable vowel shortening. Flemming (2001) is a more recent and more formally 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 duration. The exact tradeoff between elongating a syllable and compressing a segment is expressed by assigning weights to each constraint. We adopt this approach as a starting point here.

The general idea behind this approach can be captured with the metaphor of fitting malleable objects into a malleable container. Each object and the container has some inherent volume, its volume when not acted upon by external forces. 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 4](image)

**Figure 4.** Conflicting duration targets. As the number of segments inside a syllable increases, 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 other units, such as the mora, rime, foot, or prosodic word.

The auditory approach predicts a wider variety of compression patterns than the articulatory approach, due to the nature of the representations involved in the analysis. For instance, if the motivation behind segmental duration targets is to maintain auditory perceptibility of contrasts, 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 $\alpha$ masks sound $\gamma$ completely, but sound $\beta$ masks $\gamma$ only partially, the grammar may be sensitive to this distinction. Because segment duration is driven in part by auditory perceptibility, and overlap with sound $\beta$ decreases the perceptibility of $\gamma$ less than with $\alpha$, we might predict less compression of $\gamma$ adjacent to $\alpha$ than adjacent to $\beta$.

Compare this to the articulatory approach, where we would not expect compression effects to be sensitive to the auditory properties of the segments involved.

The auditory approach, then, predicts that adding consonants to a syllable will drive compression effects. It also predicts that compression might vary depending on the auditory properties of the segments involved. Sounds that mask an adjacent vowel less, or that carry more auditory information about an adjacent vowel, may license more shortening of that vowel, for instance.

2.2 Methods
2.2.1 Comparison to previous studies
The current experiment was designed to test for CS across a range of consonants and contexts in English. The study also attempts to improve upon the methodology of previous studies. Problems in that literature include a small subject pool, lack of appropriate (or any) statistical analysis, failure to distinguish between types of vowels and consonants, 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. The data were analyzed with linear mixed effects regression models, which are described in the next section. 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 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. The next section reports several methodological and statistical attempts to control for speech rate and prosodic phrasing.
The problem of segmentation criteria 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. Because segments are coarticulated, there is almost never in principle a clear point in the signal where one segment ends and the next begins. 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. Even this model is an idealization; there is often no clear point in the signal where a formant or switches 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. The objective for the segmentation strategy is to delimit intervals that are comparable across items that differ in the number of target consonants. Although the boundaries may not correspond to any psychologically real boundary between two symbols, they at least give us an acoustic landmark that can be compared across items. If we find that the interval of vowel with steady F1 in /ɡlɑ/ is shorter than that in /lɑ/, for instance, 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 /lɯ/, which track F2, or in /pʰɑ/, which track aspiration.

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.
2.2.2 Materials

The ‘target’ materials consisted of every phonotactically legal combination of the vowels \{i, a, o\} with: the consonants\{r, l, n, \emptyset\} 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. 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.

<table>
<thead>
<tr>
<th>IPA</th>
<th>Orth.</th>
<th>IPA</th>
<th>Orth.</th>
<th>IPA</th>
<th>Orth.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ad</td>
<td>od</td>
<td>od</td>
<td>oad</td>
<td>dm</td>
<td>dinn</td>
</tr>
<tr>
<td>lad</td>
<td>lod</td>
<td>lod</td>
<td>lode</td>
<td>don</td>
<td>doan</td>
</tr>
<tr>
<td>glad</td>
<td>glod</td>
<td>glod</td>
<td>glad</td>
<td>danz</td>
<td>donz</td>
</tr>
<tr>
<td>da</td>
<td>dah</td>
<td>do</td>
<td>doh</td>
<td>dnz</td>
<td>dinz</td>
</tr>
<tr>
<td>dal</td>
<td>dall</td>
<td>dol</td>
<td>dole</td>
<td>donz</td>
<td>doanze</td>
</tr>
<tr>
<td>dalb</td>
<td>dalb</td>
<td>dolb</td>
<td>dolb</td>
<td>pod</td>
<td>podd</td>
</tr>
<tr>
<td>rad</td>
<td>rodd</td>
<td>rod</td>
<td>rode</td>
<td>pid</td>
<td>pid</td>
</tr>
<tr>
<td>brad</td>
<td>brod</td>
<td>brod</td>
<td>brode</td>
<td>pod</td>
<td>poad</td>
</tr>
<tr>
<td>dar</td>
<td>dar</td>
<td>dor</td>
<td>dore</td>
<td>spad</td>
<td>spod</td>
</tr>
<tr>
<td>darb</td>
<td>darb</td>
<td>dorb</td>
<td>dorb</td>
<td>spid</td>
<td>spid</td>
</tr>
<tr>
<td>iđ</td>
<td>iđd</td>
<td>nad</td>
<td>nodd</td>
<td>spod</td>
<td>spod</td>
</tr>
<tr>
<td>lđ</td>
<td>lidd</td>
<td>nd</td>
<td>nid</td>
<td>das</td>
<td>doss</td>
</tr>
<tr>
<td>glđ</td>
<td>glid</td>
<td>nod</td>
<td>noad</td>
<td>dis</td>
<td>diss</td>
</tr>
<tr>
<td>dl</td>
<td>dil</td>
<td>snad</td>
<td>snod</td>
<td>dos</td>
<td>doase</td>
</tr>
<tr>
<td>dlb</td>
<td>dilb</td>
<td>snid</td>
<td>snid</td>
<td>dasp</td>
<td>dosp</td>
</tr>
<tr>
<td>rđ</td>
<td>ridd</td>
<td>snod</td>
<td>snoad</td>
<td>disp</td>
<td>disp</td>
</tr>
<tr>
<td>brđ</td>
<td>brid</td>
<td>dan</td>
<td>don</td>
<td>dosp</td>
<td>doasp</td>
</tr>
</tbody>
</table>

Table 1. Phonetic and orthographic representations of the 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 representations 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 before 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 nine syllables long, of the form \([X] [Y \text{ the } Z W]\), 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. 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 for a total of 360 utterances. The experiment ran between 30 and 45 minutes for all subjects.

2.2.3 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.

2.2.4 Recording
Subjects were recorded in a sound-attenuated booth inside the MIT phonetics laboratory. They were outfitted with a head-mounted condenser microphone. The utterances were recorded in mono at 44.1 kHz and saved to .aiff files.
2.2.5 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 A. For all words, the following regions were marked:

- Vowel steady-state: the portion of vowel from the innermost edge of the fixed consonant /d/ 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 /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 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.

The segmentation strategy is illustrated in figure 5.
The strength of the prosodic boundary following the target word may have varied 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 labeled 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 pitch accent on *all*, whereas the other sentences contained in the corresponding position an unaccented preposition. Most tokens of this sentence included a noticeable temporal discontinuity between the target word and *all*. We return to this irregularity in the results section.

2.2.6 Analysis

Separate models were constructed for each of two dependent variables: duration of the steady-state vowel and duration of the CV/VC transition.\(^2\) The data were analyzed with linear mixed effects regression models. 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).

Statistical analysis preceded by 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 (e.g. 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 model fit tells us how much empirical coverage we lose by strengthening the hypothesis.

The baseline model includes *random effects* of subject identity and carrier sentence. These are variables whose levels are sampled from the larger population, without covering every possible level in those populations.

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)\(^3\)

\(^2\) A third model examined the duration of consonant steady states. Those data are not reported here for reasons of space, but are consistent with compensatory shortening.

\(^3\) 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.
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, as does the duration of the fixed consonant in both onset and coda positions. Two further fixed effects are related to allophonic properties of VC words discussed in section 2.3.1.1.

The level-defining effects are vowel (\{i, a, 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, includes all 4-way interactions between these variables. Removing higher-order interactions from the model generalizes across classes of item, creating a stronger hypothesis. The statistical significance of the higher-order interactions is 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; even if these effects are non-significant, retaining them can only increase the accuracy of the estimated CS effects. After each step, MCMC simulation was repeated for the reduced model.

The significance of random effects is assessed differently in a mixed-effects model. To check for subject interactions with a fixed effect, for example, we include a by-subject random slope for the fixed effect of interest. We then compare the reduced and expanded models with a likelihood-ratio test. After the fixed effects in the model had been reduced by the procedure described above, by-subject random slopes were tested.

In what follows, fixed effects are reported with an effect size and p-value from MCMC sampling. Random effects are reported with an effect size, chi-squared statistic, and p-value from the likelihood-ratio test.

Before statistical analysis, the data were centered around 0 and normalized with 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

---

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.
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 /ðiː/, 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; most tokens fell in between these two extremes. 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 6.
Figure 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 ‘counts’ as vowel duration, both psychologically and for practical purposes. Investigating various metrics of duration is useful for what it reveals about the timing of these items and for discovering the most consistent metric to use in comparisons with other items. 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.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>i</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>58.2</td>
<td>45.1</td>
<td>39.2</td>
</tr>
<tr>
<td>m2</td>
<td>57.4</td>
<td>42.5</td>
<td>43.9</td>
</tr>
<tr>
<td>m3</td>
<td>34.3</td>
<td>36.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

Table 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 lesser 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. Comparisons within this class of item reveals something about the nature of CS. VC materials were split into five classifications of onset quality, corresponding to the tokens in figure 2.6 (a-c, e-f). The classes are 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 7 shows the average duration of each part of the VC syllables by class.

![Figure 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.](image)

The most obvious pattern here is that modal, steady-state vowel duration is longer in the classes where it is not preceded by formant transitions. This suggests that the transitions count at least partially as vowel duration in whatever sense is relevant to temporal coordination. The metric that 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 8.
Figure 8. Vowel duration across onset class and metric.

Note that, even on metric m3, duration is not perfectly uniform across classes. Creaky transition realizations have somewhat longer vowels than the others, and glide realizations have somewhat shorter vowels. One hypothesis is that transitions don’t count entirely as a part of the vowel, but that different types of transition count in different proportions. This property will figure prominently in the analysis of other phenomena in the experiment.

2.3.1.2 Comparison to CVC syllables

For all comparisons, vowel steady states in CV and VC words were significantly longer than in CVC words. This is shown below. 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/. All effects were significant below the p = 0.0001 level.

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).

To compare the size of the simplex CS effect in various contexts, we focused on items with /r/. Boundary marks between /r/ and vowels track F3 movement in all items, so landmarks are comparable across word-pairs. 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: $\chi^2(2) = 26.3; p < 0.0001$. The size of the effect differed for various subjects by up to 34 ms from the mean effect, but all subjects displayed simplex CS. Subjects also differed significantly with regard to the relative size of the simplex CS effect in onset and coda position ($\chi^2(4) = 25.9; p < 0.0001$). Three subjects showed more shortening from VC to CVC items, two showed more shortening from CV to CVC items, and one subject showed no difference between onset and coda. 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 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, $\chi^2$ 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 induce significant incremental CS in onset position (liquids: 11-15 ms; rhotics: 16-22 ms; $p < 0.01$ for both). There is even more incremental CS in coda position: 9-13 ms more for laterals, 2-3 ms more for rhotics. When the distinction between incremental CS with laterals and rhotics is collapsed, creating the class ‘liquids’ (the difference between the two is not significant), the onset-coda asymmetry is significant: 8-11 ms; $p < 0.05$. 
2.3.2.2 Nasals
Nasals induce incremental CS in onset position. It is not significantly different from the 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.

2.3.2.3 Obstruents
/p/ in /spVd/ is followed by a shorter steady-state vowel than /pʰ/ in /pʰVd/. The effect is significantly larger than the onset effect for /l/ (14-20 ms larger than /l/; p = 0.0046). The effect is reversed in coda position, producing 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.
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 (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. Vowels in VC items 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. There was a significant acclimation effect over the course of the experiment: vowels got shorter by 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. Including this variation significantly improved the fit: $\chi^2(6) = 30.2; p < 0.0001$. Some items 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.

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.

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. /pʰVd/ and /spVd/ words were excluded from this model, because their transitions (aspiration and formant transitions, respectively) are not comparable to one another. Data are shown below.
Figure 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/. The onset/coda asymmetry is significant for /o/: 8-13 ms difference, p = 0.002. It is even larger for /i/: 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 < 0.05). This averages out to a about 1-2 ms. between successive utterances of a single item.
2.3.5 Summary of results

All consonants drive simplex CS in both onset and coda position. 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:

<table>
<thead>
<tr>
<th></th>
<th>Onset</th>
<th>Coda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstruent</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Nasal</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Liquid</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 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 minimal pairs; it is not obvious what the best comparison is. While the period of steady-state vowel is shorter in /spVd/ than in /pʰVd/, /spVd/ also contains a period of (modally-voiced) formant transitions into the vowel that /pʰVd/ 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 results are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Onset</th>
<th>Coda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstruent</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Nasal</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Liquid</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 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, 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. Transitions tend to be longer in coda position than in onset position. This pattern shows idiosyncratic reversals across vowels, however, and 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 is induced by liquids in both positions, nasals in onset, and is not
clearly present for obstruent sequences in either position. Transitions between consonants and vowels do not shorten, with one exception discussed below.

In section 2.1, 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 accommodate 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 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 across the board. It also predicted asymmetries between various segments in the conditioning of CS, which would be based on the auditory properties of those segments. The auditory theory is capable of explaining most of the results, when coupled with specific hypotheses about recoverability. We briefly outline some of those hypotheses here; they will be tested empirically in the following section.

2.4.2 Recoverability and compression

The hypothesis to be explored here 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.\textsuperscript{5} 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 2.3.1.1. We saw that the particular phonetic realization of VC syllables correlates with vowel duration. Some of these tokens are produced with a glottal stop preceding the initial vowel, some with a glide transition from the vowel in preceding the. The duration of steady-state vowel is shorter in tokens that are preceded by formant transitions.

When we use 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 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 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 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 tested in the following section.

One broad asymmetry encountered here is that incremental CS is observed in comparisons of items containing liquids adjacent to the vowel, but not items containing 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 portions do in obstruent items. This is at least plausible: 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 track adjacent vowels, but we predict 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

\textsuperscript{5} 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.
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 offered here is that in some cases, the recoverability or effective duration of the adjacent vowel is at a ‘floor’ with just one adjacent 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. The idea of a minimum inherent duration is adopted from the Klatt (1979) duration model.

A further asymmetry concerns items with a nasal as C1; they show 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 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.

Liquids show a similar asymmetry to nasals, but in the opposite direction: 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 reach 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.

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 we are 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 /pʰ/ and /sp/ sequences. This study found that the steady-state vowel following /sp/ is shorter than the modally-voiced vowel following /pʰ/. 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 /pʰ/, hence induces more shortening.

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 determined by interpolation between steady-state targets, and are not manipulated (for durational properties) by the speaker. The exception is the coda /s/-/sp/ comparison, where V-s transitions shorten significantly in /Vsp/ words. Examination of the 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 three changes tend to begin closer to the same time. This suggests that, whatever the explanation of the transition shortening, it corresponds to differences in timing between the gestures that form part of /s/, rather than between consonant and vowel gestures. This intra-segment timing difference may be related to the fact, not discussed here, that the duration of /s/ itself was shorter in /dVsp/ words than in /dVs/ words. In the absence of articulatory data, we will not attempt to explain this pattern here.

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) and Shaiman (2001): 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. and Shaiman studies, and the repetition of words to a metronome beat in the Fowler study. 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, while Shaiman tests five; but these authors 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 shortening effects found in these experiments would have been meaningless if between-subject variability were taken into account.

A final possibility, noted briefly in section 2.4.2, 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.

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.
Most of the explanations of temporal patterning observed in this study concern hypothesized perceptual differences between segments or transitions. Generally, we predict that segments shorten more when the surrounding context contains more cues pertaining to their presence or quality. The next section reports an experiment that attempts to test these predictions.

3 An experimental investigation of vowel-recoverability from consonants

3.1 Introduction

3.1.1 Preliminaries

The experiment in section 2 revealed 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. We outlined a theory that can explain these asymmetries; that approach relied on several assumptions about the perceptual properties of consonants and transitions. In this section, we describe an experiment to 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 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 recoverability target of the vowel. Patterns of CS in production should be mirrored by patterns of sensitivity in perception.

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.

Based on the asymmetries in production discovered in section 2, we made 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.
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 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.

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.

We 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 /pʰ/ 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 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 driven 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.

3.1.2 Previous studies

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 we 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 discrimination) 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, and adding more vowel material back into the stimuli generates diminishing returns. It appears that the ceiling tends to occur within the first 40% of the vowel’s duration.

3.1.3 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 analyses in these papers, with the exception of Whalen’s, ignore the question of bias completely. 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 due 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 much 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. These results do not mean that /i/ is more identifiable as a category than other vowels; they mean 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 equally likely; the likelihood of correctly identifying a stimulus depends in part on a subject’s sensitivity to contrasts involving that stimulus. We can not attribute sensitivity to a category; sensitivity is a property of contrasts.

To learn about contrasts, we must construct a model that distinguishes sensitivity from bias. To 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 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. The sensitivity and bias parameters of the model were examined to confirm that they are consistent with the other experiments. Appendix B contains a detailed description of each of the analyses.

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. Subjects are more sensitive to backness/rounding contrasts than they are to height contrasts. 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, subjects reach maximum sensitivity (close to 100% correct) sometime in the first 40% of the vowel.

3.2 Methods

We constructed pairs of stimuli that differ only in their vowels: the vowel pairs tested are [e-o], [i-e], and [α-u]. The idea was to use 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 an English 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 section 2.2 and appendix A. 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 add 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 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. The figure below shows a pair of recordings used to derive stimuli for the experiment, and the segmentation strategy for creating those stimuli.

Figure 15. 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.

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,\(^6\) and three gates, for a total of 72 trials. As each sound was played, a choice of two words appeared on the screen; subjects used buttons to identify one word as the one they had heard part of. Subjects were given one 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 /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 similarity. Lexical bias, for instance, might lead subjects to respond with *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/.

For the [e-o] onset condition, 15 subjects were tested. For the [a-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 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 section 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, 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.

---

\(^6\) 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.
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. Random 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, 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 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.

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 likelihood-ratio test 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 \( \beta \), 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-square test of likelihood ratios, are reported with just the latter two values. For fixed effects, all chi-square tests have 1 Df.

3.3 Results
3.3.1 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. 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-ə] in coda and [a-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 16. 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 [a-u] onset conditions, which is significant but is averaged here for visual ease.

Figure 16 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 /pʰ/. 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/ stimuli, in particular, is not significantly above chance.

As noted above, distinctions between sensitivity parameters were removed from the model if they did not significantly contribute to the fit. The final model makes no distinction between sensitivity parameters in the [i-e] and [a-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, 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 [a-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 [a-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. In onset position before [i-e] and [a-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.\(^7\) 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 with /n/ is not significant in any condition except [e-o] onset. Here, subjects are significantly less 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 sensitivity, as well as their relative sensitivity for liquid and /sp/ items compared to other items in the experiment. Adding these differences into the model as by-subject random slopes significantly improved the fit. For overall sensitivity: $\chi^2 = 80.7, p < 0.01$. 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.

Subjects also varied in how much of an advantage stimuli with liquids had over stimuli with singleton obstruents. If we take the grand mean for this parameter as a rough guide, it suggests that 55 of the 58 subjects performed better on 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

---

\(^7\) This interaction was near-significant before by-subjects effects were added into the model (p-value of 0.03 with Bonferroni-corrected $\alpha$ of 0.0125); it was retained for this reason. After accounting for between-subject variability with regard to /p/-/sp/, however, the effect of this interaction shifted to become clearly non-significant (p > 0.05 with $\alpha = 0.0125$).
advantage for /sp/. For the [e-o] onset condition, only 8 out of 15 subjects showed the effect; 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. Incorporating this effect 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.

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.

3.3.2 Transitions
A second model examined the increase in sensitivity from adding transitions back into the zero-gate stimuli. 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 [a-u]. For /n/ in onset and coda position, none of the predicted asymmetries were observed: sensitivity to /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 with /p/.

This means that several of our hypotheses 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.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Confirmed at gate 0?</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/-/sp/</td>
<td>yes</td>
<td>as expected</td>
</tr>
<tr>
<td>/n/-obstruent, onset vs. coda</td>
<td>no</td>
<td>1 of 2 no effect, 1 opposite</td>
</tr>
<tr>
<td>liquid-obstruent, onset vs. coda</td>
<td>partially</td>
<td>1 of 2 as expected, 1 opposite</td>
</tr>
</tbody>
</table>

Table 5. Summary of results so far, showing what remains to be explained by transition data.
The figure below shows the magnitude of the transition increment associated with each type of stimulus in each condition. The figures here reflect the total increase in sensitivity from gate zero to gate two.

Figure 17. 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.

There are several things to notice about these data. Nasals show a larger transition increment in sensitivity than /p/ does in the onset conditions. In coda, however, the increment in sensitivity associated with nasals is smaller than that for singleton obstruents. Stimuli with /sp/ are associated with a larger increment in two of the three onset conditions. For liquids, we predicted that the sensitivity increment relative to singleton obstruents would be greater in coda than in onset position. This is true for the [a-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 random slopes changed the picture. With regard to fixed effects, the coda nasals pattern together; the singleton obstruents in [α-u] onset, [α-u] coda, and [e-o] coda conditions pattern together; and the /sp/ items in [α-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 [α-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 [α-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: $\beta = -2.28, \chi^2 = 26.4, p < 0.01$.

Differences between /sp/ and /p/ did not reach statistical significance in any condition.

For the comparison of liquid and singleton obstruent stimuli, patterns are mixed. In [a-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 [a-u] coda condition, leading to a significant four-way interaction between sensitivity, consonant quality, gate, and coda position: $\beta = 0.84, \chi^2 = 5.1, p = 0.02$. 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.

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: $\chi^2 = 6.1, p < 0.05$.

Subjects 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.

3.4 Discussion
The findings from this study broadly support the hypotheses put forth to explain the production asymmetries observed in section 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.

The first hypothesis holds in four of the five conditions. 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 section 2, 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.

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 [a-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 section, which are broadly similar to those for /l/.

The prediction concerning asymmetries between onset and coda liquids is supported by some findings and contradicted by others. We predicted that the transition increment in sensitivity associated with liquid stimuli was larger relative to that for singleton obstruents in coda than in onset position. This was true for [a-u] but not [e-o]. Because liquid steady states appear to contain less information about the [a-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.

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, very few of them held in [e-o] onset condition. For some of these predictions, 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 condition, or both. On the other hand, the results may simply indicate that, while the predictions of the theoretical model from section 2 hold in general, they do not hold for every single vowel contrast.

In section 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 differences in production are mirrored by perceptual asymmetries. The argument for an auditory account of compression and compensatory shortening is considerably strengthened.

4 Conclusion
This paper has described patterns of compensatory vowel shortening in English, and shown that those patterns can be explained in terms of the perceptual properties of various vowels, consonants, and transitions. Shortening phenomena differ depending on the segments present in an item, and those differences correspond to perceptual asymmetries in how much information about an adjacent vowel is carried in various consonants and transitions. Vowels shorten more when there is more information about them in the surrounding context.

One way of thinking about the logical connection between perceptual information and temporal coordination construes the conflicting pressures on duration as a tradeoff between conflicting constraints. In the current case, for instance, we could model the general phenomenon of compensatory shortening as an attempt to satisfy conflicting constraints on the target duration of a segment and a syllable (Flemming 2001). As the number of segments inside the syllable increases, these constraints will come into conflict; the conflict is resolved by some combination of segmental shortening and syllable lengthening, with the exact proportions expressed as constraint weights.

To analyze the current results, we need to posit a recoverability target for segments, which can be satisfied by both the segment proper and by information about that segment in the surrounding context; differences in the amount of vowel shortening observed adjacent to various consonants correspond to differences in the amount of vowel information contained in those consonants and their transitions. Cases where incremental vowel shortening fails to occur may be construed as cases where the vowel is at its minimum duration and can shorten no more. These correspond to cases where the amount of vowel information in the surrounding context is relatively low.

Many of the questions raised here will require extensive cross-linguistic research to be answered. These studies deal only with English; 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 contrasts 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 paper 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.
Appendix A 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).
<table>
<thead>
<tr>
<th>Segment / Context</th>
<th>V-trans</th>
<th>trans-C</th>
<th>C-C</th>
<th>C-#</th>
</tr>
</thead>
<tbody>
<tr>
<td>l/ɑ</td>
<td>F1 HP-on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l/ɪ</td>
<td>F2 HP-on</td>
<td>F2 LP-off or F1 LP-off, ER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>l/ɔ</td>
<td>F1+F2 HP-on</td>
<td>F1+F2 LP-off, ER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>l/a_</td>
<td>F1 HP-off, ED</td>
<td>F1 LP-off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>l/ɪ_</td>
<td>F2 HP-off, ED</td>
<td>F2 LP-off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r/ɑ</td>
<td></td>
<td></td>
<td>ER</td>
<td></td>
</tr>
<tr>
<td>r/ɪ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r/ɔ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r/a_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r/ɪ_</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/ɑ</td>
<td></td>
<td>F2 LP-on</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/ɪ</td>
<td></td>
<td>F1 HP-on</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/ɔ</td>
<td></td>
<td>F2 HP-on</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/a_</td>
<td></td>
<td>F2 LP-off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/ɪ_</td>
<td></td>
<td>F2 HP-off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n/o_</td>
<td></td>
<td>F2 LP-off</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pʰ/ɑ</td>
<td>Onset of energy around F1</td>
<td>Onset of aperiodic noise following burst</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pʰ/ɪ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pʰ/ɔ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p/ɑ</td>
<td>F1 HP-on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p/ɪ</td>
<td>F2 HP-on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p/ɔ</td>
<td>F1 HP-on or F2 LP-on or F2 HP-on</td>
<td>Onset of energy around F1</td>
<td>Offset of HP of energy above 5 kHz</td>
<td>--</td>
</tr>
<tr>
<td>s/ɑ_</td>
<td>F2 LP-off</td>
<td>HP of energy above 5 kHz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s/ɪ_</td>
<td>F2 HP-off</td>
<td>HP of energy above 5 kHz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s/o_</td>
<td>F2 LP-off</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d/#_</td>
<td>--</td>
<td>Onset of energy around F1</td>
<td>--</td>
<td>ED</td>
</tr>
<tr>
<td>d/_#</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɾ/ʃ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t/ʃ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- ER: End of Realization
- ED: End of Duration
Appendix B Reanalysis of previous perception studies

B.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 \(\alpha\) criterion is 0.0125. Accuracy for /i/: \(\chi^2 = 51\); \(p < 0.0001\). /u/: \(\chi^2 = 9\); \(p = 0.0022\). /a/: \(\chi^2 = 4\); \(p = 0.0351\). /o/: \(\chi^2 = 2\); \(p = 0.1792\).

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, we 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 \(i < a < o < u\).

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. 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/ (\(o, u > i, a\)).

**B.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, 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.

![Sensitivity by vowel contrast](image)

**Figure B1. 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.

**B.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, ɛ, 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 parameters of the BCM agree roughly with the other studies.
The BCM is stated as follows:

\[ p(r_j \mid s_i) = \frac{\eta_{ij} b_j}{\sum_{k} \eta_{ik} b_k} \]

It declares that the probability of response \( j \) given stimulus \( i \) \( (p(r_j \mid s_i)) \) is proportional to the similarity between \( i \) and \( j \) \( (\eta_{ij}) \) 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 \( \eta \):

\[ \eta_{ij} = \frac{Z \cdot p(r_j \mid s_i)}{b_i} = \frac{Z \cdot p(r_i \mid s_j)}{b_i} \]

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

\[ p(r_i \mid s_i) = \frac{b_i}{Z} \]

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:

\[ \eta_{ij} = \sqrt{\frac{p(r_j \mid s_i) \cdot p(r_i \mid s_j)}{p(r_i \mid s_i) \cdot p(r_j \mid s_j)}} \]

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

The BCM shows that bias is generally highest for /e/ and lowest for /e/, but patterns change by consonant. Sensitivity 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 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 B2. 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. /ɛ/). 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.

References


