# FROM INTERFACES TO SYSTEM EMBEDDING: PHONETIC CONTRASTS IN THE LEXICON

PREPRINT

**Charles Redmon***
Language and Brain Lab
University of Oxford
Oxford, UK, OX1 2JF
charles.redmon@ling-phil.ox.ac.uk

**Allard Jongman**
Department of Linguistics
University of Kansas
Lawrence, KS, USA, 66045
jongman@ku.edu

## ABSTRACT

The analysis of phonetic systems largely operates independently of other linguistic systems. This is despite the fact that the critical components of the system—phones and the contrastive relations between them—are definitionally linked to higher-order systems such as the mental lexicon. In this paper we outline an alternative approach where the phonetic system is studied as embedded within these higher-order systems. We show that system embedding radically changes the assumed role of different contrasts in the system, and warps estimates of the relative weight of the corresponding acoustic dimensions that delineate that system. Specifically, we compare three models of obstruent discrimination in English: (1) the canonical, inventory model, where contrasts are studied between phones in controlled syllables balanced in weight; (2) the lexicon model, where all contrasts are between real words; and (3) an intermediate, weighted inventory model, where the acoustics are derived from controlled syllable data but the items are sampled to match the distribution of contrasts in the lexicon. By comparing these three models we are able to identify discrepancies in the role of each acoustic dimension under different system assumptions, and model their impact on phonetic generalizations.

***Keywords*** phonetics · phonology · lexicon · interfaces · complex systems · cue integration

## 1 Introduction

The discussion of interfaces in linguistics has a long and controversial history (see Ramchand and Reiss 2007, for review). The reasons for much of the controversy are both ontological and epistemological. Ontological concerns include, for instance, questions of whether the grammar is organized into separate modules (Chomsky 1981, 2017; for general discussion of modularity, see Fodor 1983; Farmer 1985; Coltheart 1999), and whether the brain processes different subsystems of language in different regions (Fedorenko et al. 2010; Friederici 2011) or at different time intervals (Kaan 2007; Beres 2017). Epistemological disputes over interfaces largely reflect methodological differences and the domains of study assumed by different subdisciplines of the field (e.g., the sentence being in the domain of syntax, the word the domain of morphology). In phonetics, work on interfaces has largely focused on interactions between phonetic and phonological components, and connections between the theory and methodologies of their respective disciplines.

What has received less attention, however, is the relationship between phonetics/phonology and the higher-order systems such as the lexicon that ultimately determine the role of phonetic/phonological units in message encoding. This link is not only inherent in definitions of contrast and alternation, but it implies asymmetries in the role of different phonetic categories and contrasts (and the acoustic cues that subserve them) in communication that are not derivable without reference to such higher-order systems. That is, we cannot determine from independent examination of an inventory of speech sounds which sounds are most critical for listeners to detect in speech recognition, whether an $a$–$b$ target-competitor relation is equivalent to the reverse ($b$–$a$), or if one mapping between acoustic cue and phonological

*Corresponding author

feature is equal in communicative value to another (e.g., how does the mapping between vowel duration and voicing compare with that between F2 and place).

The goal of this chapter is twofold. First, we formally derive the components of the phonetic system from distinctions between higher-order units such as words in the lexicon, and argue that this derivation implies that any communicatively efficient system must consider the two systems as fundamentally interrelated. Specifically, we describe this relation as one of system embedding, where the lower-order phonetic system is seen as embedded within the higher-order lexical system. Second, we demonstrate what the consequences of this *embedded* view are for how we understand the acoustic and auditory structure of the phonetic system.

### 1.1 Prior work on phonetic interfaces

As noted above, most considerations of interfaces with phonetics, either as a component of language or as a discipline of linguistics, have focused on the phonetics-phonology interface. This focus is motivated theoretically within modular theories of language such as those in Structuralism (e.g., de Saussure 1916; Bloomfield 1926; Trubetzkoy 1939) and Generative Grammar (e.g., Chomsky and Halle 1968; Kenstowicz and Kisseberth 2014). But the focus on interfaces between phonetics and phonology is also motivated methodologically from the consideration of phonetics and phonology as two separate disciplines, each concerned with the structure and organization of speech sounds, but different in their primary methods. Historically, phonetics has utilized physical measurements in theory testing and model development; i.e., measurements of the acoustic signal, of articulator movements, of vocal tract aerodynamics, and of perception behavior. Phonology, on the other hand, has relied on naturalistic (synchronic and diachronic) observations, elicitation, and auditory impressions/intuitions.

However, as this distinction has seen much greater overlap in recent decades, most notably in work under the heading of Laboratory Phonology (Kingston and Beckman 1990), the question of where exactly the line is between phonetics and phonology has become increasingly fuzzy. For example, while anticipatory vowel nasalization was traditionally analyzed as the result of a phonological rule, nasal airflow patterns reveal that nasalization gradually increases throughout the vowel and thus should be considered the outcome of phonetic implementation rules (Cohn 1993). However, nasal coarticulation also depends on the phonology. French mid and low vowels show no nasality when preceding a nasal consonant, presumably to avoid confusion since these vowels participate in the French oral-nasal contrast (Dow 2020). Thus, the phonetic conditions that cause vowel nasalization in English are blocked by phonological constraints in French. Perspectives on the phonetics-phonology interface range from complete denial of its existence, and thereby denial of a distinction between phonetics and phonology (Ohala 1990), to clear boundaries (Hale 2000), with a range of intermediate views focused primarily on the phonetic grounding of phonology and the use of phonetic principles such as physical acoustics theory, speech aerodynamics, and articulatory/auditory physiology in explanations of phonological patterns (Kingston 2007; Scobbie 2007; see Cohn and Huffman 2014 for review).

Despite the increased overlap in approaches and methodology in the study of phonetics and phonology, one notable aspect of phonology that has largely not transferred to phonetics is the role of the lexicon. The lexicon has occupied a prominent role in phonology for many decades. This can be seen in work on the statistical distribution of phonological structures—largely under the heading of the *functional load* of phonemes, contrasts, and features—and its implications for encoding and transmission within Information theory (Martinet 1952; Hockett 1967; Surendran and Niyogi 2003, 2006). For instance, consonants generally have a higher functional load than vowels (i.e., ambiguity in consonants tends to impact information transmission more than vowel ambiguity), and among consonants distinctions in place of articulation are generally of higher functional load than manner distinctions, though languages differ considerably in the exact implementation of this asymmetry (Surendran and Niyogi 2003; Oh et al. 2015). The lexicon is also central in accounting for phonological alternations that apply differently to different subsets or strata of the lexicon, which can be seen both in the development of Lexical Phonology within rule-based frameworks (Mohanan 1982; Kiparsky 1982; Halle and Mohanan 1985) and in more recent work on the constraint-based Stratal Optimality Theory (McCarthy and Prince 1995; Rubach 1997; Bermúdez-Otero 2003, 2007).

The role of the lexicon in phonetic analysis is far more limited. Where it has received the greatest attention is in the development of models of speech recognition and speech production. However, despite the fact that many such models conceive of the mapping between the acoustic signal and the lexicon in a highly gradient fashion—e.g., the Lexical Access From Spectra (LAFS) model on the recognition side (Klatt 1979; see also Stevens 2002), the Task Dynamic model (TADA), on the production side (Saltzman and Munhall 1989; Nam et al. 2004, 2012)—development of these models has not impacted the way in which the fundamentals of the phonetic system are studied. Work in both speech production and perception still utilizes phones as its primitives (even in much of the work within the Articulatory Phonology framework, despite its commitment to alternative encodings through gestural scores), and still largely does not prioritize particular phones or contrasts within this set on the basis of their wider functional role in

the language (e.g., via their frequency in the lexicon, their onset or neighborhood densities). In the next section we provide the logical justification for this link, using the remainder of the paper to illustrate its consequences.

### 1.2 The link between phonetics, phonology, and the lexicon

The role of the lexicon in structuring the phonetic system is formally straightforward to establish. Figure 1 illustrates the derivational path from a network of lexical contrasts (a) to an inventory of phonemes (e). Here we have assumed lexical contrasts to be formally minimal pairs because that is how phonemes are typically arrived at; however, other definitions are possible. Further, this is not a circular definition, though some aspects of the derivation, such as the discrete segmentation of words, are recycled: e.g., the phoneme set, once established on a smaller subset of the lexicon, is typically applied in the description of the remainder. What is necessary then for an *embedded systems* approach to phonetics is to see this derivation as secondary. That is, the derivation may be involved in the construction of phonological grammar or in metalinguistic awareness, but the fundamental component for the organization of speech is the complex system of relations in (a).

Put more explicitly, our argument for the critical role of the lexicon in the structure of phonetic systems is as follows:

**P1.** *The primary basis of phonetic analysis is phonemic.* As much of the focus of phonetic analysis is on the acoustic and articulatory features critical to communication in a given language, phonemes are a natural unit of manipulation in the design of stimuli for production experiments, and in the analysis of identification and discrimination data from perception experiments (Pike 1972).

**P2.** *Phonemes by definition perform a lexically discriminative function.* Whether mentalist or physicalist definitions of the phoneme are adopted, a necessary condition for any phonemic encoding is that it preserves, in symbolic form, distinctions between words that are perceived by native speakers of the language to be neither auditorily nor semantically equivalent (Krámský 1974; Mugdan 1985). That is, the phoneme inventory of a language cannot be derived without some reference to the lexicon, and the points of equivalence and distinction between lexical items that allow the encoding to successfully transmit meaning through the speech signal.

**P3.** *The distribution of phonemic contrasts in the lexicon is non-uniform, with contrasts differing in their functional load.* The non-uniformity of phoneme usage in the lexicon has been documented for nearly a century **(Dewey 1923)**, and this fact has received formal phonological attention since at least 1952, when Martinet argued that the relative frequency of a given contrast was a relevant consideration (alongside articulatory and perceptual factors) in the prediction of the likelihood of that contrast to merge over time. This consideration of the relative impact of a merger as a function of the relative frequency of a contrast is referred to as the functional load (FL) of the contrast, and can be extended to distinctive features by considering classes of contrasts (Surendran and Niyogi 2003).

$\Rightarrow$ *If the speech system is optimized for transmission of phonologically encoded messages, then perceptual weighting of acoustic information in the signal must reflect this distribution.* Premises 1–3 imply the conclusion that the identification and relative weighting of acoustic cues in the speech signal must reflect to some extent the broader distribution of lexical contrasts in the language, provided that the speech system has been optimized to transmit phonologically encoded messages (taken here to be words, for simplicity). This is certainly not the only valid position—one can imagine a counter-position that the phonological inventory, though derived from lexical distinctions, exists as an independent set, and that acoustic signal parsing (and signal encoding, on the production end) operates exclusively on that set, with information flow between all higher-order units being purely phonological. But if some degree of optimization is held to apply at the level of message transmission, then any cue system structured to give equal weight to each contrast in the inventory must be sub-optimal, and therefore cue weights/rankings derived in this manner would be invalid.

In this chapter, we compare two key approaches/frameworks for the study of phonetic systems. The canonical, *independent inventory* framework is based on a balanced inventory of phones whose characteristics and relational structures are largely treated as independent of the wider system of higher-order distinctions in which they occur. By comparison, in the *lexical* framework the fundamental unit of phonetic analysis is rather a relation between contrastive words, thus shifting the state space of the system from a small, largely symmetric inventory of phones, to a large, heterogeneous ensemble of real-word contrasts in the lexicon.

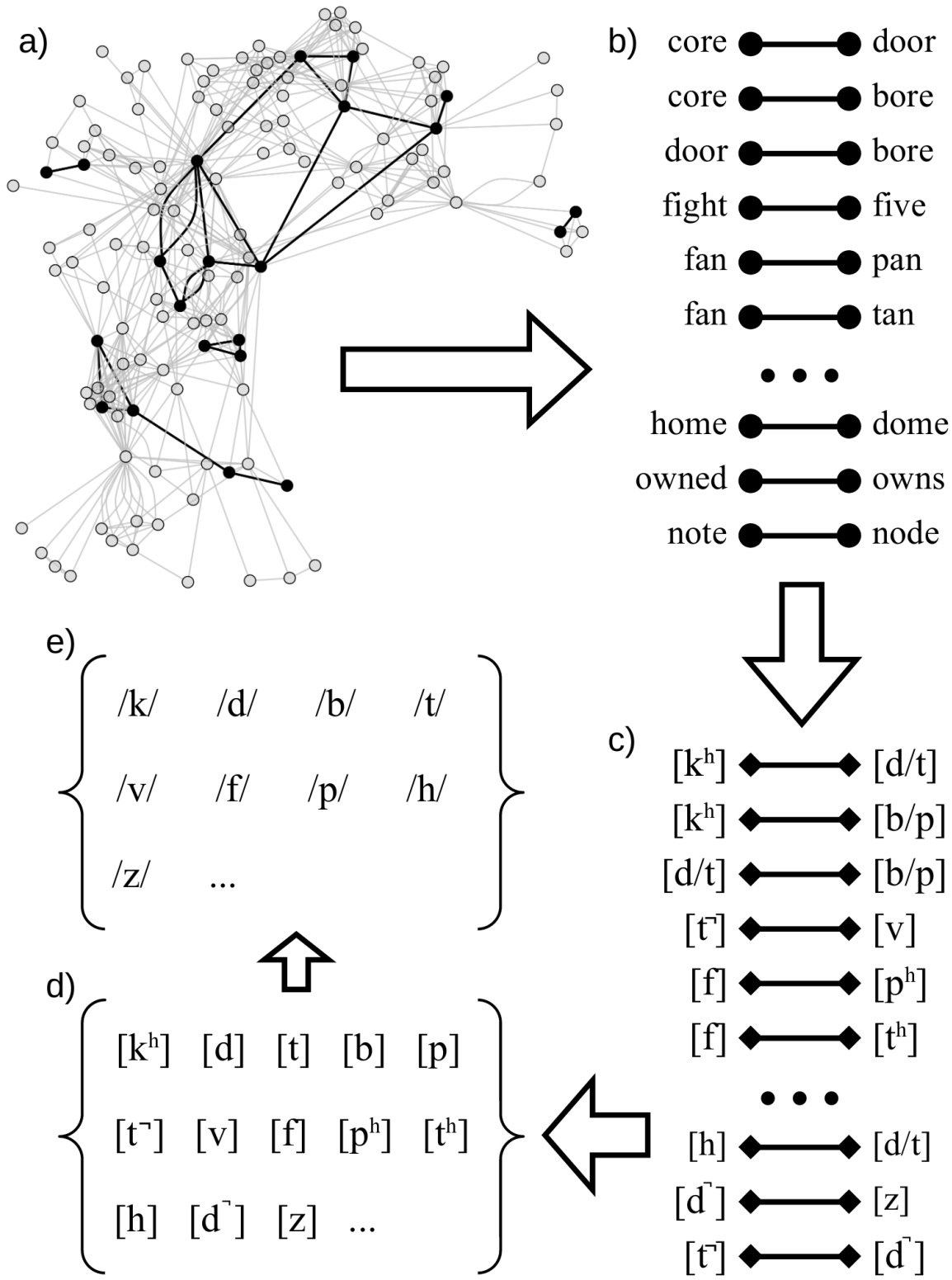


Figure 1: Illustration of the relationship between the system of minimal-pair distinctions in the lexicon (Panel *a* shows a diagram of a subset of words in the lexicon, Panel *b* lists out highlighted pairs from *a*), the set of contrastive relations between phones that define such distinctions (c), the inventory of phones comprising these relations (d), and the inventory of phonemes abstracted from the phone set (e).

### 1.3 Framework assumptions

In order to clarify the structure of the present analysis and motivate the proposed embedded system in comparison with the more canonical independent-inventory framework, we must first state several key assumptions of the two frameworks.

#### 1.3.1 Assumptions of the Independent Inventory Framework

The canonical framework applied in phonetic research treats the phone inventory as independent of the higher-order contrasts from which it derives. This framework adopts the following three key assumptions:

1. Segmentability: The acoustic signal can be segmented into discrete intervals that are used in the encoding of linguistic information.
2. Independence: The set of phones and contrasts between them exist independently of the words they encode.
3. Homogeneity: As an independent set, phones and contrasts are of equal weight in structuring the system.

These assumptions are often left unstated in phonetic analysis, model building, and experimental design, and are in some cases in conflict with the formal assumptions of the motivating theory (e.g., Exemplar theory and Articulatory Phonology do not formally operate over segmental inventories). Nevertheless, they are implicit in most phonetic research. For example, hallmark studies of the acoustic characteristics of stop consonants (Stevens and Blumstein 1978; Smits et al. 1996) and fricatives (Hughes and Halle 1956; Jongman et al. 2000) implicitly endorse these assumptions by including all voiced and voiceless phonemes in several vowel contexts that primarily yield nonwords and treating them as independent and carrying equal weight in the inventory and lexicon. This is the case even for some simulations of exemplar storage and processing (Johnson 1997b; Wedel 2006; though see Johnson 1997a for discussion on the relation between segmentation and exemplar storage, and Walsh et al. 2010 for more naturalistic, corpus-informed simulations), as well as work within Articulatory Phonology (Browman and Goldstein 1990; Byrd 1995), where the architecture of the system is based on gestural scores, but the stimulus design and organization of the analysis are often based on the manipulation of phonemic segments from an independent inventory. The point in raising these examples is to acknowledge the role inventories play even in frameworks that are explicitly attempting to move beyond canonical assumptions, and to assert that a greater centering of the *lexical* relation in the structure of the system is one way around this dependence.

#### 1.3.2 Assumptions of the Embedded Framework

When the phonetic system is considered as embedded within higher-order systems such as the lexicon, both the independence and homogeneity assumptions are eliminated. That is, the phones and contrasts are now distributed over the lexicon and thus are heterogeneous in their frequency and communicative weight. For the present study we retain Assumption 1 regarding the segmentability of the signal. This choice is made both on theoretical grounds (see, for instance, Shattuck-Hufnagel and Klatt 1979, and work on acoustic landmarks in Stevens 1992, Stevens et al. 1992, and Stevens 1998) and for simplicity as a way of isolating the contribution of the lexicon, as questions of the nature and validity of discrete representations and acoustic segmentation extend beyond the relationship between phonetic and higher-order systems. Finally, for the present study we focus only on minimal-pair relations in the lexicon, acknowledging at the outset that this is a major simplification and one that is neither necessary for the theory of embedded systems, nor correct in our view of the organization of language.

### 1.4 Present study

The present study aims to fill the gap in modeling the relationship between the phonetic system and higher-order structures by demonstrating the impact of such structures on the information encoded in each phonetic category, contrast, and feature, along with their cues in the acoustic signal. Here we focus on the English obstruent system as a test case for the implications of system embedding, though our argument extends beyond obstruents and beyond the English language, and we plan to test such extensions in future work. What we aim to show is that whether the phonetic system is considered independently of the higher-order structures it encodes, or the two are modeled as embedded systems, has far-reaching implications for the way we understand the organization of speech communication and the transmission of linguistic information in the acoustic signal.

### 1.5 Organization of the paper

The remainder of the chapter is organized as follows. In Part 2 we describe the English obstruent system and its phonological distribution in the lexicon, focusing both on the formal structure of the system and on its statistical

characteristics, the latter being essential to determining the information potential (communicative weight) for any given contrast. In Part 3 we describe the acoustic structure of obstruent contrasts and illustrate the differences between the role of acoustic cues in an independent, distribution-agnostic inventory, as compared with their role in the lexicon. This analysis serves to demonstrate the degree to which methodological isolation of lexical, phonological, and phonetic systems can bias our understanding of their structure. In Part 4 we summarize the results of several word recognition studies and their implications for the perceptual structure of the lexicon. In Part 5 we draw on English word recognition data to model the relative impact of perturbing a given acoustic dimension on listener perception. Finally, we conclude the chapter with a discussion of what asymmetries in the distribution of acoustic cues in the lexicon mean for our understanding of the structure of the English obstruent system and the analysis of phonetic systems more broadly.

## 2 English obstruent phonology and lexical contrast distributions

Before demonstrating the consequences of system embedding for our understanding of phonetic information (i.e., where the acoustic and perceptual structure of the phonetic system is integrally linked to the structure of higher-order systems such as the lexicon) we must review the formal characteristics of each system. That is, we must first identify the obstruent phones that form distinctive contrasts in English (Section 2.1), lay out their distinctive feature structure (Section 2.2), evaluate the phonotactic constraints on their composition of syllables and words (Section 2.3), and describe their statistical distribution in the lexicon (Section 2.4).

### 2.1 Obstruent phones

By most accounts there are 18 obstruents in North American English[2] (excluding minor variants such as aspirated and unreleased plosives that involve a predictable change in a single feature): six plosives (p, b, t, d, k, g), two affricates (ʧ, ʤ), nine fricatives (f, v, θ, ð, s, z, ʃ, ʒ, h), and one flap (ɾ).

Each of these phones are defined by at least one minimal-pair distinction in the lexicon, and thus represent an upper bound on the set of distinctive obstruents in the language. That is, among the obstruents in Table 1, there are a few such as the interdental fricatives (θ, ð) and the voiced post-alveolar fricative (ʒ), which are relatively low in frequency and marginal in their lexical distributions (θ/ð occurring relatively rarely and primarily in function words; ʒ occurring in a highly restricted set of French borrowings). Despite differences in theory, model architecture, and general analytical frameworks, this set is the basis for nearly all work on English obstruent phonetics, phonology, psycholinguistics, and speech engineering (see Stevens 1998, Reetz and Jongman 2020, for review).

### 2.2 Phonological feature structure

The primary purpose of featural systems is to divide the inventory into natural classes. These classes are sometimes motivated by rules governing phonological alternations, or by patterns in historical sound change, but they have also been used to account for asymmetries in speech perception (Miller and Nicely 1955; Ohala 1974; Lahiri and Reetz 2002, 2010). Several feature systems have been proposed and debated over the past several decades (Jakobson et al. 1952; Chomsky and Halle 1968; Clements and Hume 1995; Lahiri and Reetz 2002, 2010; Mielke 2008; see Halle 1990 for review), but for the present study we will adopt the classification used implicitly in most phonetic and phonological analyses. See Table 1 for the primary natural class divisions discussed in this chapter, as well as notes on classification issues and further distinctions not made in this chapter but discussed in greater detail in Redmon (2020).

In Section 2.4 we will see that the set of minimal-pair obstruent contrasts in the lexicon is not at all evenly divided among the above features, nor is each class equally represented in the constituents of such contrasts. This imbalance is consistent with prior work on the functional load of features (Surendran and Niyogi 2006), and implies that the utilization of acoustic dimensions in the lexicon will be similarly uneven.

### 2.3 Phonotactics

While phonological features provide some structure to the inventory of phones and phonemic contrasts in the language, they do not specify where in higher-order structures such as syllables or words these phones occur. In particular, the inventory is not fully free to occur in all possible combinations. There are a few key constraints on the occurrence of phones in the language, and these constraints impact any model of acoustic/perceptual information structure.

[2]Here we are using the term North American English to cover both Midwestern American English (the variety used in the perception data presented in Section 4) and Western Canadian English (the variety used in the acoustic data analyzed independently in Section 3 and used to predict listener perception in Section 5).

<table>
<tr><th></th><th>p</th><th>t</th><th>k</th><th>b</th><th>d</th><th>g</th><th>ʧ</th><th>ʤ</th><th>ɾ</th><th>f</th><th>θ</th><th>s</th><th>ʃ</th><th>v</th><th>ð</th><th>z</th><th>ʒ</th><th>h</th></tr>
<tr><td>Voicing</td><td colspan="3">–</td><td colspan="3">+</td><td>–</td><td colspan="2">+</td><td colspan="4">–</td><td colspan="4">+</td><td>–</td></tr>
<tr><td>Manner</td><td colspan="6">plosive</td><td colspan="2">affr.</td><td>flap</td><td colspan="9">fricative</td></tr>
<tr><td>Place</td><td>lab</td><td>cor</td><td>dor</td><td>lab</td><td>cor</td><td>dor</td><td colspan="3">cor</td><td>lab</td><td colspan="3">cor</td><td>lab</td><td colspan="3">cor</td><td>glot</td></tr>
<tr><td>Stridency</td><td colspan="6">–</td><td colspan="2">+</td><td colspan="3">–</td><td>+</td><td colspan="2">–</td><td>+</td><td>h</td></tr>
</table>

Table 1: Feature classes used in the present study for voicing (voiceless [–], voiced [+]), manner (plosive, affricate [affr.], flap, fricative), place (labial [lab], coronal [cor], dorsal [dor], glottal [glot]), and stridency (strident [+], non-strident [–]). Several further distinctions within these classes, such as the [high] vs. [low] coronals (distinguishing postalveolars from alveolars), the treatment of place for [h], and stridency in plosives are addressed and treated in further detail in Redmon (2020).

<table>
<tr><th></th><th>p</th><th>t</th><th>k</th><th>b</th><th>d</th><th>g</th><th>ʧ</th><th>ʤ</th><th>ɾ</th><th>f</th><th>θ</th><th>s</th><th>ʃ</th><th>v</th><th>ð</th><th>z</th><th>ʒ</th><th>h</th></tr>
<tr><td>#CV (65)</td><td>10</td><td>9</td><td>10</td><td>12</td><td>7</td><td>5</td><td>4</td><td>2</td><td>–</td><td>9</td><td>2</td><td>11</td><td>7</td><td>1</td><td>2</td><td>1</td><td>0</td><td>9</td></tr>
<tr><td>VCV (4)</td><td>7</td><td>1</td><td>16</td><td>5</td><td>3</td><td>4</td><td>5</td><td>2</td><td>17</td><td>11</td><td>2</td><td>5</td><td>4</td><td>9</td><td>3</td><td>6</td><td>1</td><td>0</td></tr>
<tr><td>VC# (31)</td><td>8</td><td>15</td><td>11</td><td>3</td><td>16</td><td>4</td><td>4</td><td>3</td><td>–</td><td>5</td><td>3</td><td>8</td><td>3</td><td>5</td><td>0</td><td>12</td><td>0</td><td>–</td></tr>
<tr><th></th><th colspan="3">voiceless</th><th colspan="2">plosive</th><th colspan="2">fricative</th><th colspan="2">affricate</th><th colspan="2">labial</th><th colspan="2">coronal</th><th colspan="2">dorsal</th><th colspan="3">strident</th></tr>
<tr><td>#CV</td><td colspan="3">69.7</td><td colspan="2">51.9</td><td colspan="2">41.4</td><td colspan="2">6.6</td><td colspan="2">31.6</td><td colspan="2">44.7</td><td colspan="2">14.6</td><td colspan="3">24.8</td></tr>
<tr><td>VCV</td><td colspan="3">50.5</td><td colspan="2">35.3</td><td colspan="2">40.2</td><td colspan="2">7.6</td><td colspan="2">32.1</td><td colspan="2">47.9</td><td colspan="2">20.1</td><td colspan="3">22.8</td></tr>
<tr><td>VC#</td><td colspan="3">56.7</td><td colspan="2">56.2</td><td colspan="2">36.0</td><td colspan="2">7.8</td><td colspan="2">19.7</td><td colspan="2">65.7</td><td colspan="2">14.7</td><td colspan="3">30.7</td></tr>
</table>

Table 2: Percentages of phone and feature class occurrences in minimal-pair contrasts in a representative subset of the English lexicon (see Section 3.1.1 for a complete description of the model lexicon and the *Lex95* set used in these calculations). Percentages of the total set of contrasts that occur in each position are indicated in parentheses in the first column.

At syllable onset all obstruents are licensed, though the alveolar flap may only occur intervocalically before an unstressed syllable, and the voiced postalveolar fricative /ʒ/ is highly restricted in onset position (limited to only a few French borrowings). As a corollary to the alveolar flap distribution, the alveolar plosives from which it derives are generally not found in VCV position, except at the onset of a stressed syllable or in cases of context-driven hyperarticulation. By comparison, syllable codas are more constrained, with [h, ɾ] not licensed in VC position. Word-finally, obstruents are even more constrained, though this is not reflected in absolute constraints so much as in probabilistic ones, with word onsets much more evenly distributed among different obstruent phones and featural classes than word offsets. See Table 2 for a breakdown of phone and feature probabilities at word onset (#CV), intervocalically (VCV), and at word offset (VC#).

The asymmetries in Table 2 have several implications for speech perception. For example, the greater density of contrasts (higher phone entropy) in #CV position means that cues to the consonant that occur at the beginning of the following vowel have the potential to play an outsized role in the system. Vowel offset cues, on the other hand, should be relatively more marginal, with the exception of those vocalic cues that allow for the discrimination of the highly frequent coronal contrasts (namely among [t, d, s, z]) word-finally.

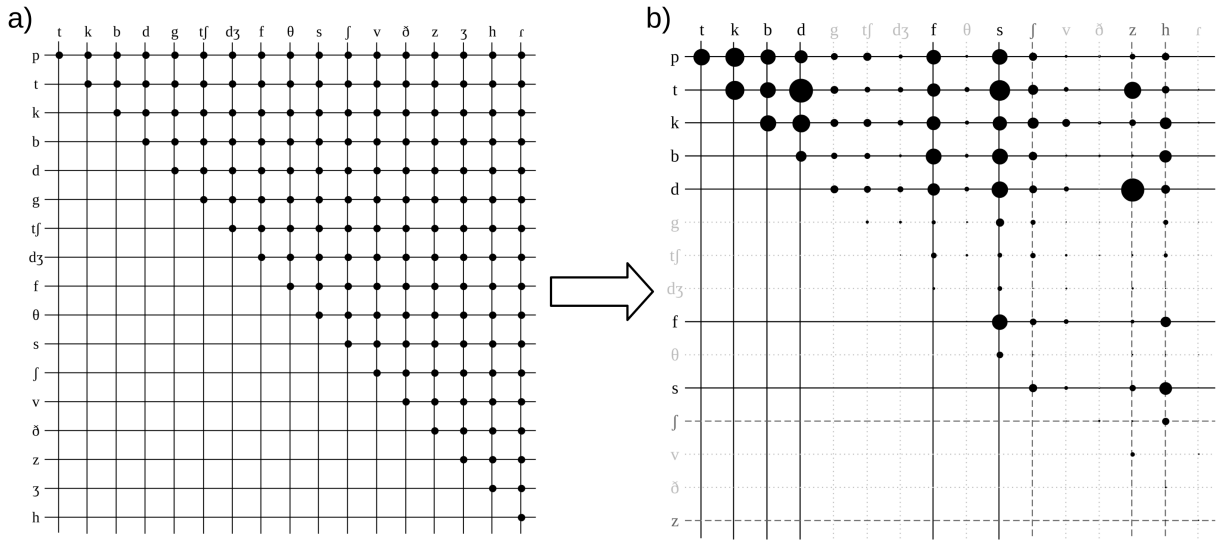


Figure 2: Illustration of the warping of phonetic inventory structure due to asymmetries in the lexicon. On the left (a), we observe the lattice structure assumed in the inventory framework, while on the right (b), the majority of contrasts involve a limited subset of the inventory: primarily the plosives [p, t, k, b, d] and the voiceless fricatives [s] and [f].

### 2.4 Lexical contrast distributions

Finally, in order to understand the formal implications of a shift from an independent to an embedded system, we must account for the statistical distribution of phones, contrasts, and features in the lexicon. This account does not necessitate the direct usage of statistical information in speech production and perception, but it does provide context for the relative frequency at which an English listener must utilize a particular cue in speech recognition, as well as how often they must manipulate these acoustic dimensions in their productions.

Figure 2 illustrates how conditioning on the lexicon leads to a warping of the phone and contrast weights from a uniform lattice in the case of the independent inventory, to a lexical system that is heavily reliant on approximately a third of that system. In particular, the contrasts among plosives and between plosives and [s], [f], and to a lesser extent [h] and [z], with the latter primarily operative in word-final contrasts.

## 3 Acoustic structure of independent and embedded systems of contrast

The central point of comparison in the present study is between the pursuit of phonetics as an independent system, and one which is embedded within the higher-order systems it derives from. This distinction has both theoretical and methodological implications, the chief methodological consequence being the common study of acoustic phonetics as a function of balanced, controlled syllable sequences, either from nonwords or real-word exemplars. An embedded approach, by comparison, requires the use of a large database of real words to approximate the complex system of lexical distinctions in the language. Thus for example, instead of focusing on the acoustic characteristics of [s] and [ʃ] in English, and quantifying the relative perceptual weight of different cues through a series of experiments on s–ʃ discrimination (perhaps using a range of techniques, such as behavioral tasks, eye tracking, or EEG), we might ask where in the lexicon does the s–ʃ contrast serve to distinguish meaningful units, what other contrasts are present in the system and relevant to the word recognition process, and ultimately how do these two sibilants, and their acoustic/perceptual characteristics, serve in the encoding and decoding of linguistic messages. This study aims to test the implications of that shift at three levels: acoustically (how do the acoustics of real-word contrasts differ from controlled syllables, and what implications do such differences have for models of cue structure), perceptually (how are cues predictive of listener word recognition behavior, and do syllable-based inventory models scale to real-word perception), and at the level of the organization of the system (how can we simulate the role of different cues and contrasts in maintaining distinctions in the lexicon, and ultimately in preserving the transmission of information). Below we review the different sources of acoustic data we used to model the fundamental operations of each framework.

### 3.1 Data

Several data sources are used for the present analysis, including both controlled syllable and real-word data, which can be divided into two classes: single-speaker and reference data. The single-speaker data, which is used as stimuli in the perception experiments (Section 4) and as the basis for simulations of cue perturbation (Section 5), comprises approximately 27,000 words and 1,000 nonword syllables produced by a single speaker: a 28-year-old male native speaker of Western Canadian English. Thus, single-speaker data focuses on within-speaker characteristics of the phonetic system. Reference data integrates multiple existing databases of controlled syllables that have served as stimuli in past perception experiments, and may therefore be used as a baseline for the cue weighting models of the inventory system.

#### 3.1.1 Single-speaker data

All single-speaker data were recorded as part of the Massive Auditory Lexical Decision project (Tucker et al. 2019), and for the purposes of this study have been divided into two parts: a *model lexicon* and a *model inventory*.

**Model lexicon.** A sample of 26,793 words was compiled by Tucker and colleagues from several sources: all unique word types in the Buckeye corpus (∼8,000; Pitt et al. 2007), an additional 9,000+ words from the English Lexicon Project (Balota et al. 2007), the 10,000 next highest frequency words in the Corpus of Contemporary American English (COCA; Davies 2009), and 1,252 compound words from CELEX (Baayen et al. 1995). This corpus was then subdivided into two acoustic data sets that will be used in the analysis below. The first set, referred to as the *Recognition Set*, is a set of 960 obstruent-contrastive minimal pairs that were used in the perception experiments introduced in Section 4. The second set, referred to as the *Lex95* Set, is a subset of the MALD database, used for the system perturbation simulations in Section 5, that includes the most frequent words comprising 95% of the tokens in each of several corpora—COCA (written and spoken; Davies 2009), SUBTLEX-US (Brysbaert et al. 2012), and Google Unigram (Michel et al. 2011). This criterion reduces the 26,793-word MALD database to a core subset of 3,406 words participating in 11,972 minimal-pair contrasts. Within this set 2,501 minimal pairs (from 1,649 words) involve contrasts between obstruents.

**Model inventory.** A total of 1,020 controlled syllables exhibiting obstruents in initial (CV), final (VC), and medial (VCV) positions were recorded from the same speaker who produced the stimuli in Tucker et al. (2019). These items were recorded to facilitate direct comparison between controlled syllables and real words while holding the speaker constant. CV items were of the form: [$CV$b], where $C$ is the set of all permissible obstruents in word-onset position—namely, [p, t, k, b, d, g, ʧ, ʤ, f, θ, s, ʃ, h, v, ð, z, ʒ]—and $V$ is the set of monophthongal vowels [i, ɪ, e, ɛ, æ, ʌ, ɑ, o, ʊ, u], yielding 170 items × 2 repetitions = 340 CV syllables. VC items were similarly constructed with [b] as the onset consonant, 10 monophthongal nucleus vowels, and 16 offset consonants ([h] is excluded from the above set as it is phonotactically illicit), yielding 320 VC syllables (160 items × 2 repetitions). Finally, VCV sequences were constructed with the form [b$V_1CV_2$b], were $V_1$ is the same monophthongal vowel set, $C$ is the 17-obstruent consonant set, and $V_2$ was constrained to match $V_1$ itemwise (i.e., only symmetric vowel contexts were recorded). In addition to the 340 items generated from this template (17 consonants × 10 vowels × 2 repetitions), 20 items were added to elicit the alveolar flap, [ɾ]. Such items were of the form: [ˈb$V$ɾəb], with $V$ varied between the 10 monophthongs and stress always on the first syllable, and were repeated twice.

#### 3.1.2 Reference data

The following corpora were used in the construction of a reference data set. For controlled syllables, data from the California Syllable Test (CaST; Woods et al. 2010) were used. All controlled syllables were recorded from phonetically trained speakers. For the CaST, four speakers of Midwestern American English (2 female, 2 male) were recorded producing 20 onset consonants in 3 vowel contexts [i, a, u] with 20 coda consonants. This 9600-syllable database was then reduced to a reference set for the present study consisting of 360 items that closely match the form of the CV and VC sets in the target inventory data: 192 [$CV$b] items (16 onset obstruents × 3 vowels contexts × 4 speakers × 1 repetition) and 180 [b$VC$] items (15 coda obstruents × 3 vowels × 4 speakers × 1 repetition).

For the VCV reference data, we utilize the test stimuli from the 2008 Consonant Challenge organized at Interspeech 2008 (CC08; Cooke and Scharenborg 2008), which comprises 384 items chosen for presentation to listeners from much larger set of 12,096 recordings (24 consonants × 9 vowel contexts × 2 stress types × 28 speakers). The 9 vowel contexts were formed from every possible combination of the corner vowels [i, a, u]. That is, unlike in the target data, vowel contexts were not symmetric in the reference data. Further, Cooke and Scharenborg did not record the alveolar flap [ɾ] as part of their study, which is both a part of the inventory data recorded from the target speaker and a critical part of the lexical contrast system in English. These characteristics make the CC08 database a less compatible reference for intervocalic obstruent contrasts than is the CaST for CV and VC contrasts, but given the shortage of VCV

databases with accompanying perception data we must simply acknowledge this discrepancy as a possible confound. From the test set, which contains 2 productions of each consonant by each of 8 speakers (4 female) for a total of 384 items (though items are not controlled to match in vowel contexts), we arrive at 272 items with obstruent consonants (17 obstruents $\times$ 2 items $\times$ 8 speakers), which were further reduced to a final reference set of 254 items after removing repetitions.

### 3.2 Acoustic cue definition

A wide array of acoustic measurements have been proposed to distinguish various subsets of the English obstruent system, typically, though not exclusively, focused on delineating categories along a single featural dimension (e.g., voicing, place, sibilance, etc.). This choice, while often for experimental control, clarity of argument, or simply convenience has in some instances constrained cue definitions to be undefined for certain classes of sounds (e.g., voice onset time, VOT, is canonically defined exclusively for plosives).[3] Further, such constraints have led to inconsistencies between the theoretical coverage of a given cue (the range of contexts where that cue may be identified in the signal, based on its acoustic definition) and its experimental coverage (the contexts tested in the literature).

In the present study we focus on 22 acoustic cues, which can broadly be classified into temporal (7), amplitudinal (3), and spectral (12) measurements. There are several cues beyond this set (e.g., the spectral moments of Forrest et al. 1988 and Jongman et al. 2000, and the Discrete Cosine Transform, DCT, coefficients of Bukmaier and Harrington 2016) that have been excluded for two reasons: (1) to limit the number of highly correlated parameters (e.g., Spectral Peak Frequency and Spectral Center of Gravity), and (2) to focus only on those parameters that have clearly identifiable articulatory sources (DCT coefficients in this sense are more global measurements reflecting a composite of underlying causes). In Table 3 we list each cue within the temporal, amplitudinal, and spectral sets, alongside publications where they were first identified or notably extended.

A key point of note in considering the above cues is that this set arose entirely out of considerations of independent inventory discrimination. However, given that each of these cues was independently motivated from physiological considerations, both articulatory and auditory, we have reason to believe that the set used in word recognition should not be entirely different. What remains an open question, however, is the relative utility of each cue when applied to the general problem of word recognition.

### 3.3 Modeling acoustic discrimination

From the cues listed above we next ask the question of how such cues may be optimally integrated to distinguish between meaningful linguistic units. Here the primary focus is on the utility of each cue in the lexicon, and the degree to which such cue weights are predictable from an independent inventory of contrasts. The independent inventory framework has two major features that we will show play an outsized role in the results: (1) the units being balanced means that equal weight is applied to cues distinguishing contrasts and featural dimensions of different functional loads; (2) the acoustics are more uniform and often hyperarticulated (even if they come from real words they come from a controlled subset), though we will see that this is not always to the benefit of contrast.

The models described in this section may be described as *ideal perceiver* models, in that they model contrast discrimination under the assumption of perfect recognition. In the ideal perceiver model, contrast presence, $y$, is a dichotomous variable (0 = same, 1 = different), and each acoustic cue in the predictor matrix $\boldsymbol{X}$ is included as the absolute difference between the cue value in one member of the contrast and the cue value in the other; i.e., for the contrast between *bit* and *pit*, $\Delta\text{VOT} = |\text{VOT}_{bit} - \text{VOT}_{pit}|$. These parameters are referred to in the remainder of the chapter as *contrast parameters*. The statistical model employed in mapping the predictor set $\boldsymbol{X}$ onto the outcome $y$ is a Bayesian Additive Regression Tree, BART (Chipman et al. 2010), which was chosen for two reasons. First, as a tree-based model it is relatively less sensitive to multicollinearity in the predictors than are linear models such as logistic regression and linear discriminant analysis. Second, the ensemble learning method used in BART, where predictions are derived from the aggregation of several simpler models, reduces bias from any one model in a manner similar to the Random Forest algorithm, but with fewer problems with overfitting given the Bayesian method of model fitting based on weakly informative priors. Finally, separate models were fit to CV, VCV, and VC contrasts, where the parameters included in each model are only those which are defined for that position. We should note that this is not a model of spoken word recognition, but rather a model of the information in the acoustic signal that distinguishes any given obstruent-contrastive minimal pair, and thus this is a model of cue *potential*, where cue weights are used as an

[3] Here we use the term "cue" rather than "parameter" for clarity in later presentation of model parameters in Section 3.3, though we recognize that a distinction is often made between cues as *percepts* and parameters as *acoustic measurements* that may or may not directly correspond to the perceived input.

| **Temporal Cues** | **Spectral Cues** |
|---|---|
| Consonant Duration ($DUR_C$) [7, 13, 23] | Spectral Peak Frequency ($FREQ_{PK}$) [2, 8, 14, 15] |
| Vowel Duration ($DUR_{V1/V2}$) [1, 3, 7] | Spectral Peak Amplitude ($AMP_{PK}$) [4, 25, 32] |
| Closure Duration (CD) [10] | Dynamic Amplitude ($AMP_{DYN}$) [31] |
| Noise Duration (ND) [7, 8, 9] | Spectral Tilt ($TILT_{C/V1/V2}$) [26, 32] |
| Voice Onset Time (VOT) [16, 18] | Spectral Shape (SHAPE) [33] |
| Voice Cessation Time (VCT) [29] | Spectral Dispersion ($DISP_{C/VC/CV}$) [21, 24, 34] |
| Voicing Percentage (VOI%) [29] | Low Frequency Energy (LF) [8, 35] |
| | Relative Amplitude of F3 ($AMP_{F3}$) [27, 30] |
| **Amplitudinal Cues** | Relative Amplitude of F5 ($AMP_{F5}$) [27, 30] |
| Burst Presence (BURST) [12, 22] | Fundamental Frequency ($f0_{CV/VC}$) [3, 19] |
| Noise Amplitude ($AMP_N$) [14, 15] | First Formant Frequency ($F1_{CV/VC}$) [5, 6] |
| Vowel Amplitude ($AMP_{V1/V2}$) [11] | Second Formant Frequency ($F2_{CV/VC/V1/V2}$) [5, 6, 28] |
| | Third Formant Frequency ($F3_{CV/VC}$) [17, 20, 32] |

Table 3: Temporal, amplitudinal, and spectral cues used in the present study. See the following references for initial cue definitions and notable later uses: [1] Lloyd (1936), [2] Liberman et al. (1952), [3] House and Fairbanks (1953), [4] Fischer-Jørgensen (1954), [5] Liberman et al. (1954), [6] Delattre et al. (1955), [7] Denes (1955), [8] Hughes and Halle (1956), [9] Gerstman (1957), [10] Lisker (1957), [11] Lehiste and Peterson (1959), [12] Fant (1960), [13] Peterson and Lehiste (1960), [14] Strevens (1960), [15] Heinz and Stevens (1961), [16] Lisker and Abramson (1964), [17] Öhman (1966), [18] Lisker and Abramson (1967), [19] Haggard et al. (1970), [20] Fant (1973), [21] Gray and Markel (1974), [22] Dorman et al. (1977), [23] Umeda (1977), [24] Stevens and Blumstein (1978), [25] Ohde and Stevens (1983), [26] Lahiri et al. (1984), [27] Stevens (1985), [28] Krull (1989), [29] Docherty (1992), [30] Hedrick and Ohde (1993), [31] Shadle and Mair (1996), [32] Smits et al. (1996), [33] Evers et al. (1998), [34] Jongman et al. (2000), [35] McMurray and Jongman (2011).

indicator of which cues are potentially most useful in distinguishing items in the lexicon. Whether and to what extent this cue structure is predictive of listener behavior is a question addressed in the next section.

In what we refer to as the inventory model, the input matrix, $\boldsymbol{X}$, is the set of acoustic parameters from controlled syllable productions from the single-speaker data, which includes only contrast parameters in the ideal perceiver model (later, in the listener model we will use both contrast parameters and the direct cue values of the target item, referred to as *target parameters*). The response, $y$, is the contrast presence/absence (1/0) vector where contrasts were created by systematically pairing each obstruent in the 17-phone set (18 in VCV), controlling for position and vowel context, and equivalence relations (within-category comparisons) were created by pairing the two repetitions of the same item.

The lexicon model aims to predict the presence/absence of contrast in the *Lex95* set directly from the acoustic characteristics of the words participating in such contrasts. One important theoretical feature of this model is that since the acoustic measurements are made from the minimal pairs themselves, and the outcome predicted by the model is based on contrasts rather than specific obstruent categories, the lexicon model operates on a relaxed assumption regarding the phonological composition of each word, as it only knows that two intervals in a word pair are contrastive (i.e., are acoustically distinct in a linguistically meaningful way) and that they fit within the broad featural class of [–sonorant]. Within-category comparisons are then made in the lexicon model by pairing words with shared onset, middle, or offset syllables (depending on the CV, VCV, or VC position of the contrast).

Finally, in addition to the inventory and lexicon models, we further consider a *weighted inventory model* wherein both contrast and vowel-context distributions are scaled to match those in the lexicon, though the acoustics are derived from controlled syllable data. In other words, in the weighted inventory model the fundamental acoustic properties of English obstruents are held to be independent of their phonological distribution in the lexicon. With this assumption, however, we can still study how the lexicon might warp cue weights based on which information is more likely to play a role in listeners' perception of distinctions between real words. This is done by sampling the acoustic measurements from controlled syllables—i.e., drawing observations from the input matrix $\boldsymbol{X}$ in the inventory model—such that the proportion of occurrence of each obstruent contrast in a given position and vowel context matches that in the lexicon.

From patterns of agreement and disagreement between the three models (lexicon, inventory, and weighted inventory) we are able to describe both the general scalability of acoustic cues between different architectures of the phonetic system—i.e., between one that assumes an independent inventory of speech sounds and one that is embedded in the ensemble of form distinctions in the lexicon—and the source of discrepancies where they arise; e.g., is a cue upweighted or downweighted in the inventory relative to its role in the lexicon because of fundamental differences in the acoustics of the two data sets, or because of differences in contrast distributions that fail to emphasize the characteristics of the contrasts which play the greatest role in the lexicon?

Next we examine each of the three main classes of model relations: agreement, distributional disagreement, and acoustic disagreement. The remaining cases where estimated cue weights do not scale between the two frameworks may be classified as a composite of distributional and acoustic discrepancies (see Redmon (2020) for further details and examples).

### 3.4 Points of agreement in lexicon and inventory models

While the focus of this chapter is to illustrate the degree to which the assumptions made by the independent-inventory framework bias our estimates of the wider role of different acoustic cues in the language, it is important to demonstrate cases where the two frameworks nevertheless agree. What must be determined in such cases is whether these points of agreement are due to stable relationships between the articulatory and auditory systems that have driven the organization of the system (as in Stevens 1972), or if they are accidental (i.e., did we get lucky as researchers in studying an aspect of the system that is similarly acoustically and phonologically distributed in the inventory and lexicon?). Likely the answer is some combination of the two.

Figure 3 shows a case of cue agreement in vowel-onset F2 ($\text{F2}_{\text{CV}}$). Figure 3a plots the relationship between vowel-onset F2 distinctions ($\Delta\text{F2}_{\text{CV}} = |\text{F2}_i - \text{F2}_j|$, for contrast pair $i, j$) among contrasts in controlled syllables in the inventory model ($x$-axis) and among real-word contrasts in the lexicon model ($y$-axis). For example, the *b–d* pair in Figure 3a shows that on average there is an approximate 300 Hz difference in F2 onset between *b*- and *d*-initial syllables and words (the $x$ and $y$ axes, respectively), whereas for the k–ʃ contrast the F2 onset distinction is slightly larger among real words than among controlled syllables. Figure 3b shows the relative change in predicted contrast likelihood (the partial dependence functions) for $\text{F2}_{\text{CV}}$ in the inventory, weighted inventory, and lexicon models. From Figure 3 we see the source of the close agreement between the three models. Not only do both the inventory and lexicon exhibit a wide range of $\Delta\text{F2}_{\text{CV}}$ values in contrastive pairs, but the majority of this range lies above the 75th percentile of the within-category values in both data sets.

Examining Figure 3 we find clear evidence of vowel-onset F2 indexing place of articulation, both within and across manner of articulation. At the upper end of the $\Delta\text{F2}_{\text{CV}}$ range, showing minimal pairs with a relatively large difference in F2, are frequently occurring contrasts such as *b–d*, *d–f*, and *d–h*, along with several contrasts involving the voiceless sibilant fricatives [s, ʃ]. In fact, one of the most robust distinctions is that between postalveolar and non-postalveolar obstruents, as the set [ʃ, ʒ, ʧ, ʤ] exhibits consistent F2 onsets in the 1.5-2 kHz range that are well above the onset F2 frequencies of most other obstruents. At the lower end of the $\Delta\text{F2}_{\text{CV}}$ range, showing minimal pairs with a relatively small difference in F2, are contrasts such as *b–p*, *b–f*, and *f–p* that overlap largely with the within-category range due to their shared place of articulation (other voicing contrasts such as *k–g* and *t–d* are more distinct but are not labeled in Figure 3 because they are relatively less frequent). Place contrasts between voiceless obstruents (excepting the aforementioned contrasts involving sibilants) are also relatively less distinct in F2 onsets compared to voiced contrasts or contrasts between voiced and voiceless obstruents (due to $\text{F2}_{\text{CV}}$ partially reflecting voicing differences). For this reason $\text{F2}_{\text{CV}}$ is a robust predictor of contrastiveness, whether adopting a lexical or inventory framework.

### 3.5 Model disagreement due to misalignment of phonetic distributions in the inventory and lexicon

Misalignment of cue weights between the two frameworks commonly occurs when there are major distributional differences in the occurrence of contrasts that depend on a particular acoustic dimension for their discrimination. One such case is preceding vowel duration ($\text{DUR}_{\text{V1}}$) among word-final contrasts. The role of $\text{DUR}_{\text{V1}}$ is substantial in

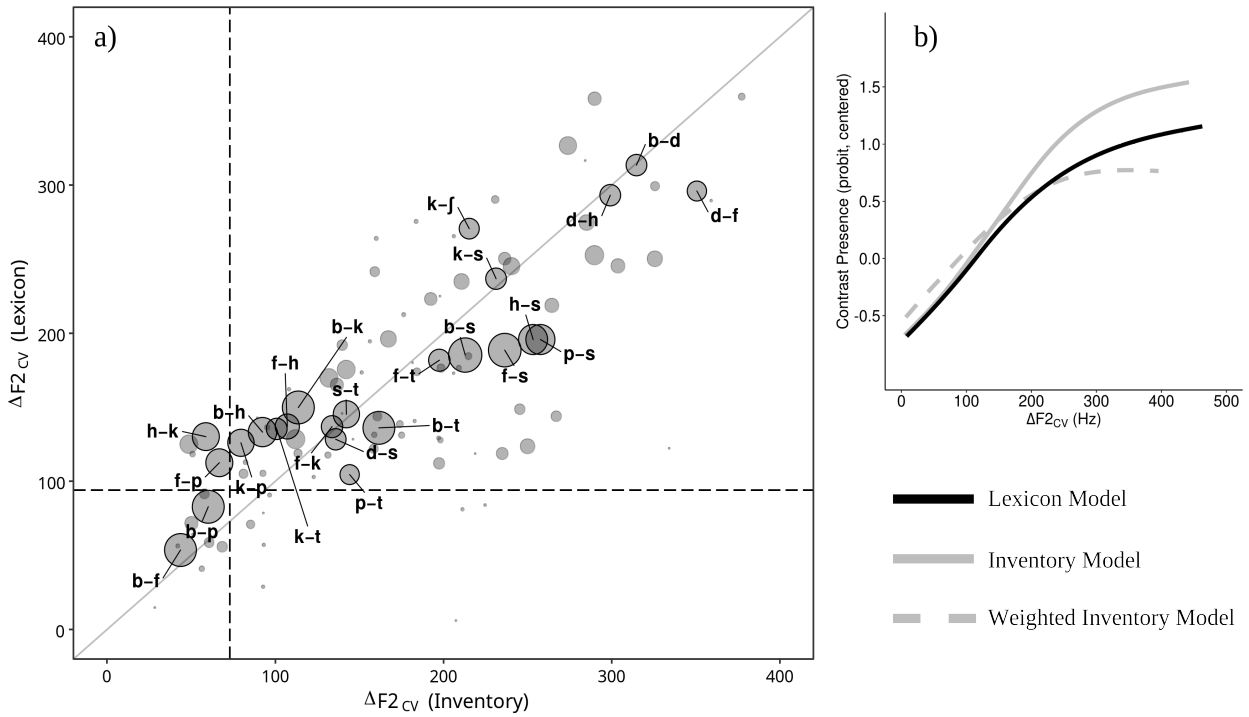


Figure 3: A case of cue agreement between the two frameworks. Word-initially there is a close correlation between $F2_{CV}$ distinctions for different phonetic contrasts in the inventory and lexicon models. Dashed lines indicate the 75th percentile of the within-category range. Point size is scaled to match lexical frequency. Figure (b) shows the partial dependence functions of the likelihood of contrast presence (relative to its absence in within-category distinctions) as a function of differences in F2 onset in each model. Note: contrasts not present in the lexicon are excluded from the plot since they cannot be matched with the inventory measurements.

all three models, with a moderate discrepancy in cue ranking between the inventory and lexicon. Nevertheless, by accounting for the distribution of items in the lexicon this distinction is largely eliminated as the weighted inventory model is brought into complete agreement with the lexicon in terms of aggregate cue weight on $DUR_{V1}$. Figure 4 shows that among the most frequent contrasts, *d–z* and *t–z* (and to a lesser extent *d–k*), vowel duration differences are slightly larger in the inventory than in the lexicon, meaning that when such contrasts are given greater weight in the weighted inventory model the end result is an increase in the distinction between contrastive and non-contrastive vowel duration differences relative to the inventory model. More critically for the interpretation of vowel duration as a cue to word-final voicing, however, is the relatively diminished role of within-voicing contrasts in the lexicon, as with the exception of *d–z*, most such contrasts are well within the range of within-category variation and thus offer no role for $DUR_{V1}$ as a cue for discrimination. Since these contrasts are equally represented in the inventory the overall role of preceding vowel duration is downweighted relative to its role in the lexicon.

Other examples of distributional disagreement between the inventory and lexicon models include cases like the above, where cue weights are overestimated in the inventory, such as noise amplitude word-initially, F3 onset/offset word-medially, and vowel-offset F1 word-finally. Opposing cases, where cue utility is underestimated due primarily to differences in distributional assumptions within the two frameworks, include low-frequency energy word-initially, and spectral peak amplitude and spectral dispersion word-medially. These cases illustrate that there are acoustic features whose role in an embedded lexical system can be estimated from the statistical distribution of contrasts in the lexicon, but their role when studied in a balanced inventory may be under- or over-estimated. Such cases are problematic for phonetic research more broadly because the balanced inventory approach is what the majority of present and prior research is based on.

### 3.6 Poor scaling between inventory and lexicon due to acoustic differences in the data

We turn finally to poor scaling between the inventory and lexicon due to acoustic discrepancies between contrasts as they occur in controlled syllables and those present in real-word distinctions. Noise amplitude among word-final

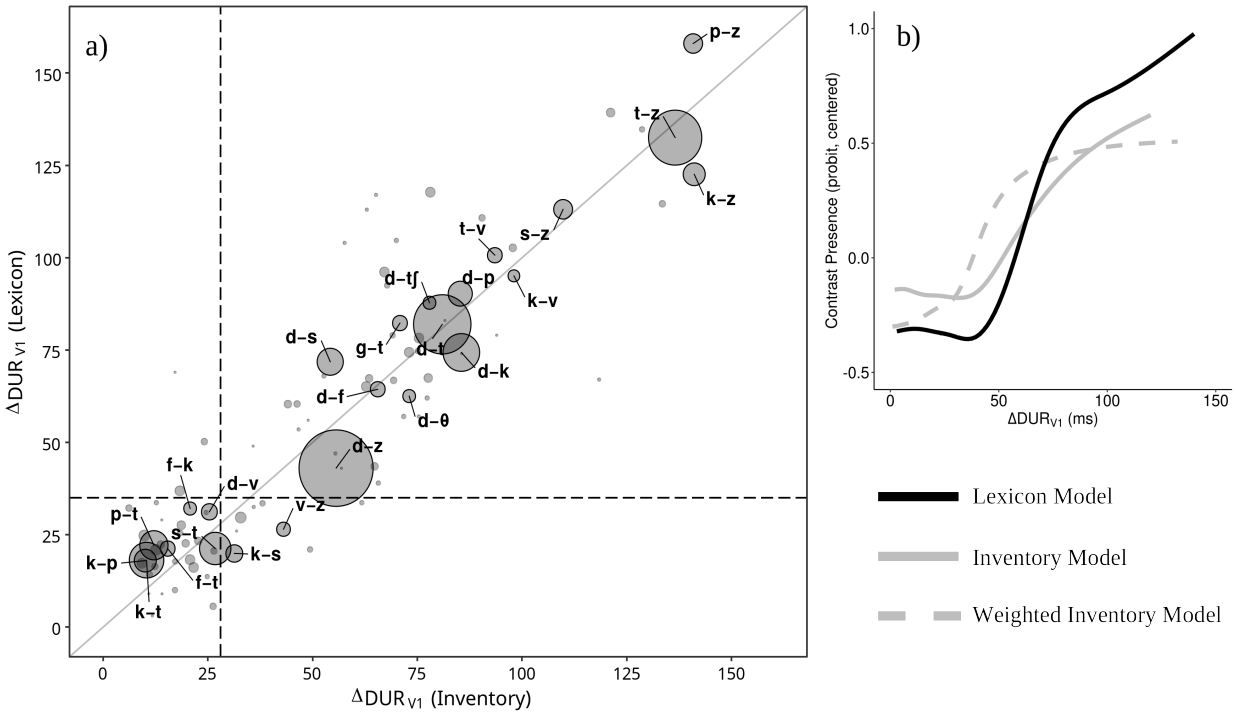


Figure 4: A case of distributional disagreement between the two frameworks. Word-finally there is a close correlation between $\mathrm{DUR_{V1}}$ distinctions for the most frequent contrasts in the lexicon, but there is a large cluster of contrasts with similar preceding vowel durations (making vowel duration less useful as a cue to those contrasts) that appear rarely in the lexicon, meaning the lexicon model is free to weight $\mathrm{DUR_{V1}}$ more highly. Note also that while the lower end and steepness of the weighted inventory curve in (b) is more similar to the lexicon model than the balanced inventory curve (the solid gray line), the threshold has shifted to a shorter duration due to the gap in $\mathrm{DUR_{V1}}$ distinctions along the $x$-axis (30-50 ms) that is absent in the lexicon.

contrasts is one such cue. Figure 5 shows the relationship between noise amplitude distinctions in the controlled syllable data and in the real-word data, and illustrates that noise amplitude is poorly correlated between the two, though it is not immediately clear how the distribution of $\Delta\mathrm{AMP_N}$ values among phonetic contrasts results in a relative upweighting of noise amplitude in the weighted inventory model. The majority of the contrasts in the lexicon are more distinct (i.e., above the gray identity line) in the lexicon than in the inventory, meaning that the discriminative power of noise amplitude in the weighted inventory should be reduced rather than enhanced. The only explanation then is that the change in cue weights between the inventory and weighted inventory is due rather to the reduction of within-category variance from the controlled syllable data, which creates a lower boundary between contrastive and non-contrastive items. This shift, combined with the narrowing of the contrastive range (note there is much greater variance along the y-axis than the x-axis), results in a more rapid jump in contrast likelihood around 5–7 dB in comparison with the more gradient shift in the lexicon.

Cases such as these motivate the need for models to operate on real words rather than simulated analogues built on combinatoric manipulations of nonword syllables. There are many other such cases in the ideal perceiver models. Cases like that in Figure 5 of an overestimate of cue utility in the weighted inventory include VOT and vowel-onset F1 in word-initial position, noise duration and low-frequency energy word-medially, and voice cessation time (VCT) word-finally. Alternatively, there are cases of acoustic misalignment leading to underestimates of lexical cue utility, such as noise burst presence and consonant voicing percentage word-initially, spectral tilt and VCT word-medially, and spectral peak amplitude and dispersion word-finally.

In analyzing cue integration within an ideal perceiver framework, we were able to explore the encoding potential of a wide range of cues when optimally weighted for the discrimination of obstruent consonants within a given system. Across positions, the most influential cues in the lexicon were F2 at vowel onset/offset, consonant voicing percentage, consonantal spectral tilt, and spectral peak frequency. Many of these cues were also highly weighted in the inventory model, while cues such as noise duration and spectral shape were more consistently upweighted relative to the lexicon

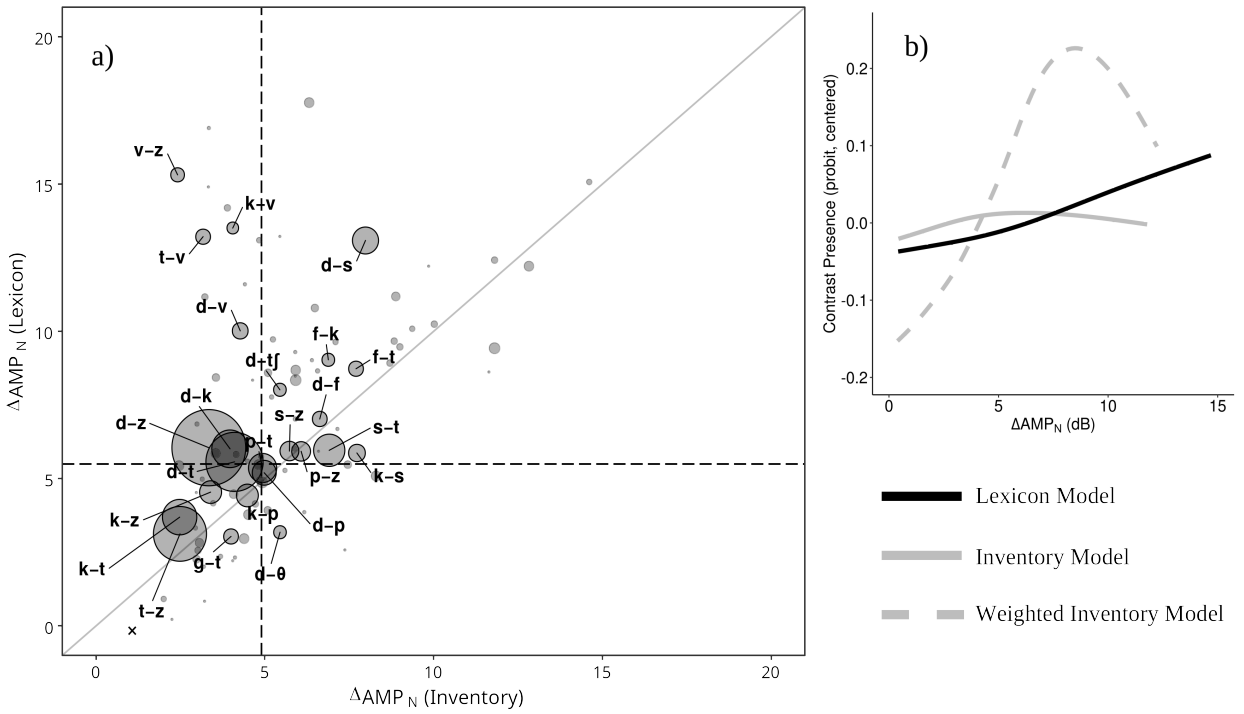


Figure 5: A case of acoustic disagreement between the two frameworks. Word-finally, Figure 5a shows that obstruent noise amplitude ($AMP_N$) is moderately more distinctive between contrasts in real words relative to controlled syllables (perhaps due to the equalizing effects of controlled productions), as contrasts are more widely distributed along the $y$-axis than the $x$-axis. However, because the acoustics are misaligned between the two data sets (note that there is almost no correlation between noise amplitude in nonword syllables versus real words), merely scaling the syllable estimates to match the lexical distribution results in an erroneous overestimate of cue strength in the weighted inventory model (Figure 5b).

model. Vocalic cues, on the other hand, particularly f0, F3, and vowel-midpoint F2, were generally of little utility, either in the lexicon or inventory systems.

This taxonomy of sources of cue disagreement serves primarily to account for how the behavior of the system depends on both acoustic and phonological distributions, and to provide some means of linking the inventory and lexicon models in an explanatory way. This procedure is critical to understanding how the inventory assumption impacts our understanding of cue integration, and how previous findings from the canonical study of phonetic systems in the literature can be incorporated into this new lexical framework in the future. However, one problem with the ideal perceiver models is that they can only distinguish between contrastive and noncontrastive items. Distinctions in acoustic salience within the contrastive set can only be distinguished by incorporating data from listener perception, which we examine in the next section.

## 4 Perception of lexical contrasts as compared with controlled syllables

From Section 3 we can see several acoustic dimensions along which real-word distinctions differ from controlled syllable sequences, and therefore where ideal perceiver models predictably vary in their relative cue weights. The next question to address then is whether this difference is reflected in differences in auditory sensitivity as well, and here again is an area where inventory and lexical frameworks are expected to differ notably, because in recognizing real words listeners can be expected to utilize top-down biases, such as word frequency/familiarity biases. In this section we use the same acoustic data to predict listener recognition behavior, and thereby arrive at estimates of cue weights that more closely reflect the integration of acoustic information in speech perception. Further, one of the challenges noted in Section 3 was the fact that in many instances the model was unable to distinguish between contrasts that are well beyond the within- vs. cross-category threshold along a particular cue dimension. This is because the model only categorizes pairs as either contrastive (1) or non-contrastive (0), meaning that neither formal (feature distance)

nor perceptual (auditory salience) distinctions are captured in the model. However, the listener models, in tracking the relative discriminability of each contrast, can capture these differences and thus provide more gradient information on the role of each cue in the system.

### 4.1 Listener data

As with the acoustic data, listener data were divided into two groups: those responding to real words from the single-speaker recordings, and those responding to controlled syllable data from the reference sets. The former refer to word recognition data collected in the present study (Section 4.1.1), and the latter to data from previous studies on consonant recognition in controlled syllables (Section 4.1.2).

#### 4.1.1 Word recognition data

Eighty native speakers of American English responded to 960 obstruent-contrastive minimal pairs (split into two separate lists of 480 distinct minimal pairs each) in a two-alternative forced choice (2AFC) task. For example, listeners heard *butter*, after which they were given two options to select from: *butter* and *buffer*. Items were evenly divided in the position of this contrast between word-initial (CV), word-medial (VCV), and word-final (VC) contexts, and were otherwise distributed between mono-, di-, and tri-syllabic lengths approximating that observed in the lexicon. For each session half of the stimuli were presented in six-talker babble noise at –2 dB SNR, and half at +2 dB SNR. These SNRs were chosen based on a pilot experiment targeting between 70 and 80% accuracy overall, and an approximate 10% difference in accuracies between the two SNRs. Where otherwise unstated, model results are presented for the aggregate behavior on both SNRs.

#### 4.1.2 Syllable recognition data

Reference perception data for the inventory models was drawn from Woods et al. (2010) for the CV and VC models, and Cooke and Scharenborg (2008) for the VCV model. The results reported below focus on the mapping between the target speaker and the reference listener data in order to hold the speaker constant when evaluating the three models; however, parallel models were run with the reference acoustic data as input to confirm that both configurations result in a similar fit (see Redmon 2020 for details).

### 4.2 Model structure

The structure of the listener models is largely the same as in the ideal perceiver models, with two main differences. First, in the cue set the contrast parameters used in Section 3 are supplemented with the corresponding direct parameters from the target word in each minimal pair—referred to hereafter as the *target parameters*—as well as the absolute frequency of the target word and the frequency of the target relative to the competitor. For example, the ideal perceiver model in Section 3 uses $\Delta$VOT to detect whether *bit–pit* is contrastive (as compared with *bit–bit*), whereas the listener model presented here uses both the VOT of *bit* and the relative difference between *bit* and *pit* ($\Delta$VOT), as well as the frequency of occurrence of *bit* and the relative difference in frequency between *bit* and *pit*, to predict listeners' mean accuracy in recognizing *bit* when *pit* is given as a competitor. Word frequency is not included in the parameter set for the inventory model, as inventory-based cue weights are assumed to be independent of higher-order information such as that from word frequency. Second, the response $y$ in the listener model is the probit-transformed listener accuracy on each item. All other aspects of the model are the same as before.

### 4.3 Effect of incorporating listener sensitivity in lexical and inventory models of cue integration

In shifting to models of listener perception, we recall the important observation in McMurray and Jongman (2011), among others, that tracking human performance, rather than optimal performance, is the ultimate goal in work on human speech communication. In the present study this means accounting for the constraints of perception in noise, where we predict the background babble noise to least impact cues at the extreme upper and lower ends of the frequency range, as well as durational cues that do not depend on low-amplitude feature detection (i.e., vowel duration should be more robust than VOT). Figures 6 and 7 illustrate the shift in cue rankings between the ideal perceiver and the listener models of word-initial and word-final contrasts, respectively. This mapping is further divided into relative cue ranks in the lexicon and inventory.

Overall, we find greater distinctions (a lower correlation) between the inventory and lexicon cue ranks in the listener model than in the ideal perceiver model, which makes sense given the additional factors involved in predicting listener recognition behavior (characteristics of the background noise, lexical biases, individual differences in auditory acuity). Word-initially, the cues in closest agreement between the inventory and lexicon models of listener recognition

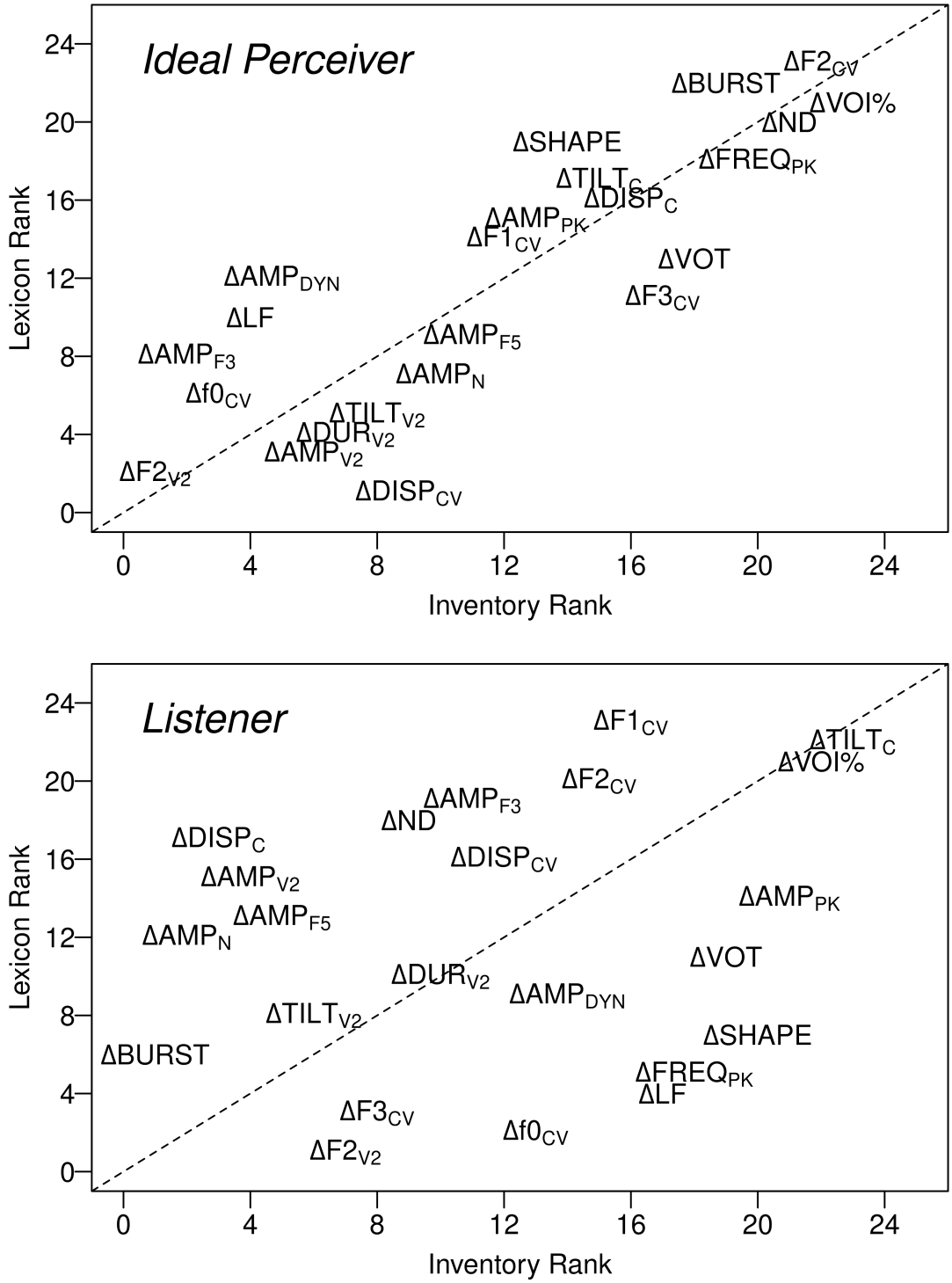


Figure 6: Relation between cue rankings in ideal perceiver and listener models of word-initial (CV) obstruent discrimination. The top panel shows the results for the ideal perceiver models; the bottom panel shows the results for the listener models. The dashed line on the diagonal indicates points of ranking equivalence.

(the bottom panel of Figure 6) are consonant voicing percentage (VOI%) and spectral tilt ($TILT_C$), as they are both consistently implemented, serve to distinguish a dense array of contrasts (voicing in the former, place and sibilance in the latter), and are perceptually salient.

When it comes to points of disagreement between the inventory and the lexicon, there are several cues that diverge between the inventory and lexicon models, whether for an ideal perceiver or for human listeners. For example, Figure 6 shows that the roles of both VOT and F3 onset in the lexicon are overestimated by the inventory model, and that this is consistent for ideal perceiver and listener models (though F3 onset is much less critical overall in the listener model than in the ideal perceiver model). This implies that whether or not listener perception is taken into account, VOT and F3 are reliably less informative among real-word contrasts than in a balanced inventory of controlled syllables. This result is perhaps not surprising because both cues are heightened in single-feature distinctions (voicing among plosives for VOT; place between dorsal and coronal plosives), and such distinctions are greatly enhanced in a balanced inventory relative to the asymmetric lexicon which relies much more on multi-feature distinctions. Further, both cues are correlated with cues such as F1 and F2 onset, respectively, that are more versatile (serving in a variety of featural distinctions among obstruents) and perceptually reliable.

In terms of discrepancies between the listener and ideal perceiver models, there are several notable cases that illustrate the importance of incorporating listener perception into models of cue structure and integration. First, consider the cues spectral peak frequency ($FREQ_{PK}$) and spectral shape (SHAPE), which are both highly underutilized in the lexicon model once listener perception is accounted for, but in the ideal perceiver model they remain critical in both systems. Both of these cues rely on spectral characteristics in the mid-frequency range (for most obstruents) that may become distorted when listening in the presence of background noise. Further, both can be recovered (to infer place of articulation, for example) from F2, which is not only more robust acoustically given that in the vowel it is excited by an open, resonating voice source, but its predictable trajectory into the middle of the vowel makes interpolation of F2 onset easier for the listener than static cues in the consonant noise. Note further that F2 has a considerably greater weight in the lexicon in the listener models than in the inventory.

Turning to cues that are underutilized in the inventory relative to the lexicon, but only when listener perception is taken into account, the relative amplitude in the F3 region is one such cue. $F3_{AMP}$ is one of the highest ranked cues in predicting listener word recognition behavior in the lexicon, while it is a mid-range cue in the inventory (see the bottom panel of Figure 6, where $F3_{AMP}$ is well above the identity line in the listener models). However, $F3_{AMP}$ ranks low in both ideal perceiver models (lexicon and inventory). Relative F3 amplitude is a cue primarily to sibilance distinctions, both between fricatives (the case it was first proposed for) and in fricative-plosive contrasts, making it valuable in the lexicon where word-initial obstruent contrasts are dominated by the plosives and the fricatives *s* and *f*. In an ideal perceiver model there are many other cues, such as burst presence/absence (BURST) and spectral shape (SHAPE), that offer similar discriminatory information for these contrasts, but as noted earlier these cues do not appear to be easily detected by listeners. Relative F3 amplitude, on the other hand, as a transitional cue at the CV boundary is more recoverable as its information is more distributed.

These are just a few cases of discrepancies in word-initial cue integration between the inventory and lexicon approaches which are highlighted by listener perception data on heterogeneous word sets versus balanced syllables. Below we review examples from word-final position, which is marked by even greater asymmetries in obstruent distribution and perception.

Among word-final contrasts we see the same lower correlation among inventory and lexicon cue ranks in the listener model relative to the ideal perceiver model. Both assign high weights to noise duration, a cue that consistently indexes several features, most notably manner and voicing, and indicate that listeners weight preceding vowel duration more highly than is assumed under ideal recognition (both in the inventory and lexicon, though the shift is greater in the lexicon). Vowel duration is a critical cue to word-final voicing distinctions in general, but from model comparisons the mapping of vowel durations in controlled syllables aligns poorly with that expected for real-word distinctions (note too that this result cannot derive from simple *absolute* shifts in duration in different speech contexts, but requires changes in *relative* temporal relationships between different contrasts).

Word finally, the greatest discrepancies between ideal perception and listener behavior are in the use of consonant voicing percentage (VOI%) and preceding vowel F2 ($F2_{V1}$). The former can be explained by the relative rarity of word-final obstruent voicing in English (at least in isolated word production where such cases are also in utterance-final position). In real-word production in particular, phonologically voiced and voiceless obstruents do not differ substantially in the degree of voicing in the consonant interval; rather, they differ in the duration and formant dynamics of the preceding vowel. Nevertheless, these differences are consistent in the acoustics and thus usable in the ideal perceiver model; they are just not as reliably detected in perception, particularly in the presence of background noise.

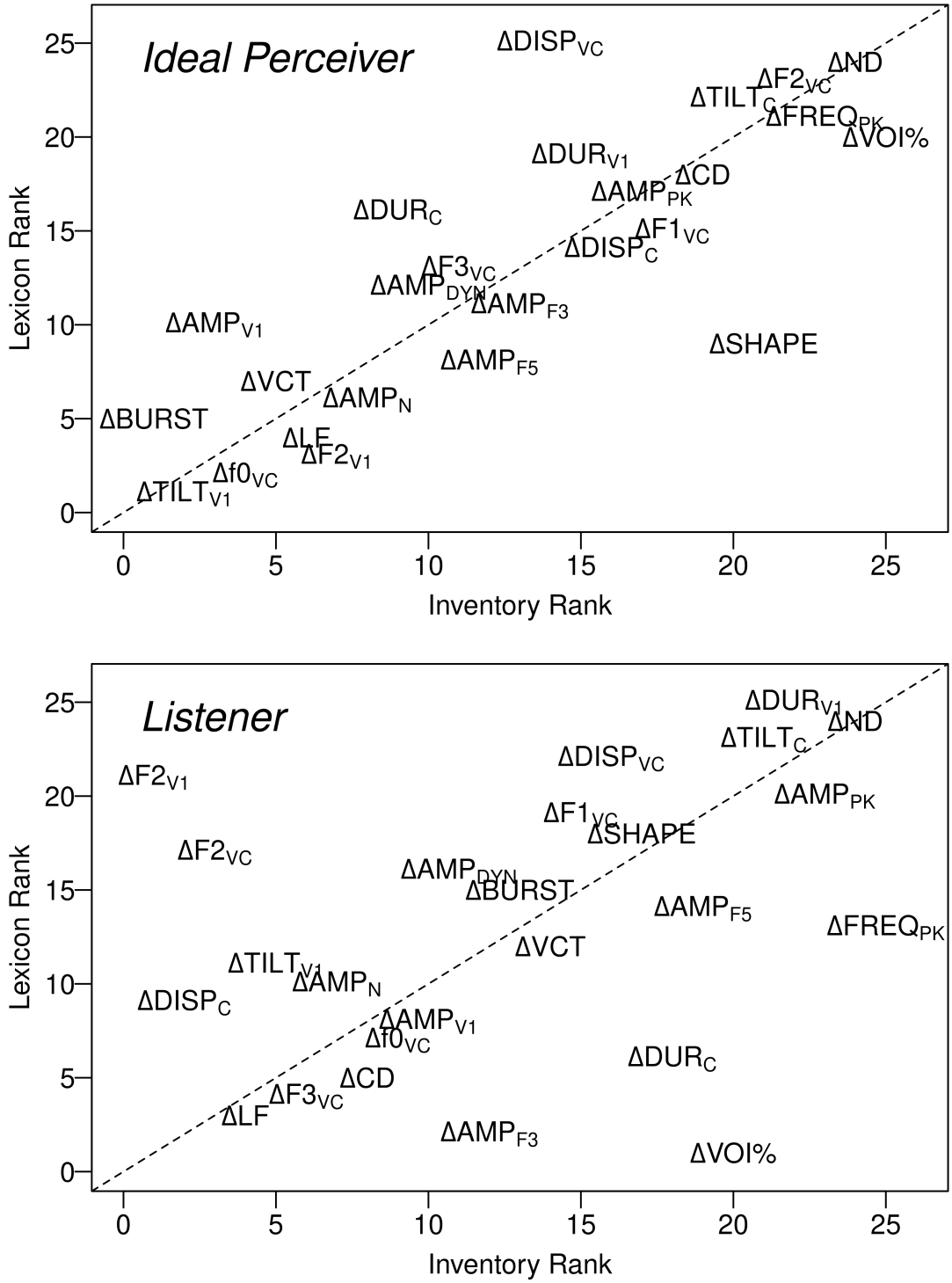


Figure 7: Relation between cue rankings in ideal perceiver and listener models of word-final (VC) obstruent discrimination. The top panel shows the results for the ideal perceiver models; the bottom panel shows the results for the listener models. The dashed line on the diagonal indicates points of ranking equivalence.

Regarding $F2_{V1}$, we have noted earlier that coarticulation is both greater and more informative in word recognition than in controlled syllable perception. By spreading consonantal place of articulation information well into the preceding vowel (not just at vowel offset, but at vowel midpoint as well), listeners can often detect the identity of the consonant even if cues in the noise interval are obscured by background noise. This fact has been understood since Öhman (1966), but it remains obscured when production and perception data focuses on more controlled contexts where coarticulation is reduced and where the availability of other correlated cues (such as spectral characteristics of obstruent noise) is greater than would be observed in more ecological conditions.

### 4.4 On word recognition more broadly

The examination of cue integration in the prediction of listener recognition behavior was motivated primarily on the basis of improved validity as a measure of cue utility in communication over an ideal perceiver framework with no mechanism to account for relative differences in obstruent contrast perceptibility. However, this section also introduces an important methodological point in the scaling problem between inventory-based estimates of cue weighting and cue weighting in spoken word recognition. The inventory model operates on both controlled syllable acoustics and perception; and indeed, much of the work on speech perception in the phonetic literature relies on perception data from either balanced syllable stimuli or a small sample of real words meant to exemplify a particular subset of the phonetic system. Given that in addition to acoustic and distributional discrepancies, there were many discrepancies in cue weights attributable to perception differences between the two experiments, this raises the question as to whether listeners' cue parsing behavior is at all sensitive to the item and choice constraints of the task.

Clearly, listeners deploy a much wider array of top-down information in word recognition than the simple model of word frequency biases that we were able to apply. Further, we have not been able to address the temporal dynamics of lexical access, and thus cannot account for cohort competition effects or temporary uncertainty that is resolved later when more of the signal is processed. These and many other naturalistic constraints on word recognition are a necessary part of a full account of phonetics as an embedded system; indeed, with higher-order systems as central to the structure of the phonetic system the role of these factors is more readily apparent.

## 5 Lexical consequences of acoustic and auditory perturbation

In making the above acoustic and perceptual comparisons we were able to locate both points of agreement and disagreement between the two frameworks, and where disagreements arose we identified the primary source of the discrepancy in differences in the distribution of obstruent contrasts in the lexicon and inventory or differences in the acoustics/perception of real words versus controlled syllable sequences. This analysis was primarily listener-centric in studying both the information available to listeners, and how they appear to be integrating that information in perception based on their word recognition behavior.

This section may be described as lexicon/system-centric in that it examines the impact on the system of contrasts in the lexicon of the relative discriminability of items as a function of two sources of uncertainty: (1) background noise, and (2) cue loss. More broadly, we use the simulation of contrast weakening in the lexicon under acoustic perturbation to study how an obstruent system that is embedded in higher-order distinctions is structured. This analysis provides both a new way of studying functional load that is gradient—i.e., it reflects the relative likelihood that the information contained in obstruent contrasts in the lexicon is preserved in the presence of background noise—and one that operates below the phone, phoneme, or feature by examining the degree to which a given cue contributes to the maintenance of lexical form distinctions. Both analyses are critical to an approach to acoustic phonetics that is fundamentally linked to the higher-order systems that are encoded in the signal.

In the sections that follow we simulate both general auditory perturbation (5.1) and cue perturbation (5.2) by assigning weights to lexical contrasts in the *Lex95* network that reflect the relative confusability (predicted error rate) of the contrast. Figure 8 shows, for example, how the effect of weakening particular contrasts through noise or cue degradation depends on how the affected contrasts are distributed across the lexicon.

### 5.1 Auditory perturbation

We simulate auditory perturbation by examining the impact of background noise in the stimulus on the separation of obstruent contrasts in the lexicon. Since we do not have direct access to listener responses to all obstruent-contrastive minimal pairs in the Lex95 database (due to constraints on item repetition in word-recognition experiments) the relative discriminability of such items in the model lexicon was estimated from the listener word recognition data in Section 4. In particular, a linear regression predicting probit-transformed accuracy from Target and Competitor obstruents ($C_T$, $C_C$), contrast Position (CV, VCV, VC), Noise Level (+2 dB SNR, –2 dB SNR), Word Length (mono-, di-, tri-syllabic),

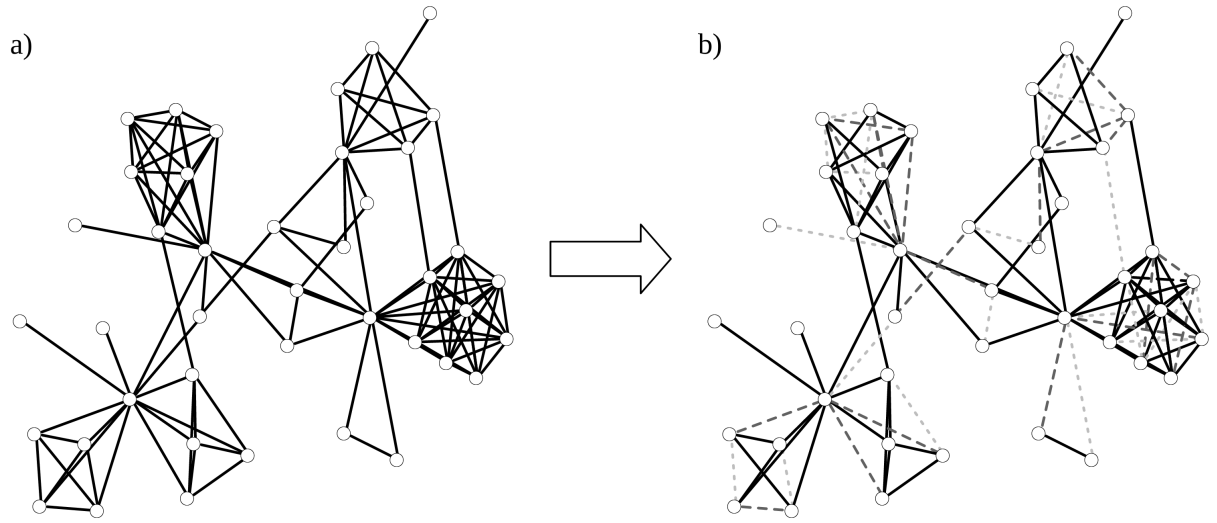


Figure 8: Illustration of the consequences of lexical contrast perturbation: either general auditory perturbation (as in the addition of background noise to speech stimuli), or cue perturbation (as in the selective manipulation of particular cues or acoustic dimensions). Here (a) is the original sample network of lexical contrasts, and (b) is the perturbed network, where contrasts that are weakened are shown in gray dashed and dotted lines. As contrasts are dependent on lexical entries, the effect of any perturbation is assumed to be a complex function of lexical structure.

Absolute Target Frequency (the log frequency of the target), and Relative Target Frequency (the log frequency of the target minus the competitor) was run on the data. Overall, this model provides a close fit to the data ($R^2$ = 0.49), which crucially provides contrast accuracy predictions whose rank distribution correlates highly with the observed values, both overall ($\rho$ = 0.991) and by Position ($\rho$ = 0.993), Noise Level ($\rho$ = 0.996), and Word Length ($\rho$ = 0.885). That is, from this combination of predictors we are able to closely replicate the general pattern of contrast discrimination in the recognition set (the items that were presented to listeners in the 2AFC task), allowing us to then project those predictions on a more complete lexicon model in the form of the Lex95 database, which we then use to study the impact of background noise at both the higher (+2 dB) and lower (–2 dB) SNRs.

Overall, the predicted impact of signal perturbation by background noise (the mean impact of the +2 dB and –2 dB SNRs relative to a no-noise baseline) on the number of minimal pairs comprising the obstruent system in the lexicon was an 18% reduction in expected minimal pair count (i.e., summing the minimal pair discrimination probabilities, where the ideal case of perfect discrimination then reduces to the regular phonological minimal pair count): from 2501.0 to 2055.5, consistent with the error rates designed for the study based on pilot data. Further, the effect of a 4 dB reduction in SNR was a reduction in predicted minimal pair counts of approximately 11% (from 2176.0 minimal pairs at +2 dB to 1935.1 minimal pairs at –2 dB), again consistent with the general effect of SNR on listener recognition. These results are not trivial, because the Lex95 data is largely distinct from the stimulus set in the 2AFC experiment, and also includes words of a much higher baseline frequency (again, based on its design as a lexical core representing only those words which comprise most of everyday English usage), which could have biased the predictions toward higher accuracies, though this outcome was not obtained.

In addition to measuring perturbation effects through minimal pair counts, we examine the change in the functional load of phones and contrasts in the lexicon due to background noise. Functional load not only treats minimal pair counts logarithmically, but it can be extended to a gradient measure of contrast weight through the following equation from Surendran and Niyogi (2003): $FL(x) = \sum_y P(x,y)FL(x,y)$, where $P(x,y)$ equals the predicted error rate of listeners on a given contrast, and $FL(x,y)$ is the change in entropy (computed over word units) from a merger of $x$ and $y$. In the case of single-phone analyses, functional load is measured by taking the sum of the resulting weighted FL measure over the set of contrasts that phone occurs in. Thus for example, if the functional load of f–θis 1 and the functional load of f–s is 5, and listeners have a 30% error rate on f–θ, while they only misidentify f-s 5% of the time, then the weighted functional loads of f–θand f-s are 0.3 (0.3 × 1) and 0.25 (0.05 × 1), respectively, and the functional load of f is 0.55 (0.3 + 0.25). This may seem a bit counterintuitive, because we may think of f-s as more *functional* in the system as it is both more frequent and more reliably perceived, but here what we are measuring is rather the relative *criticality* of a contrast: its potential to cause information loss. Similarly, a contrast like ʃ–n may occur in

many lexical oppositions in English, but it is less critical to the speech production and perception systems because it is highly unlikely to be misperceived or misarticulated. Finally, the adoption of the word rather than the segment as a unit of measurement in the calculation of entropy changes makes *FL* and *minimal pair count* similar, but functional load is still a valuable addition because of its greater generality and extensibility to non-minimal-pair contrasts in future work.

Figure 9 shows the aggregate change in minimal pair counts and functional load for contrasts involving each obstruent phone as a function of noise level (+2 dB SNR, –2 dB SNR). From Figure 9 we see that the most robust class are the sibilants, at a 20% reduction from the baseline phonological counts to the expected counts at –2 dB, while flaps and affricates are the least robust at 26 and 25% reductions, respectively. Within each feature the relation between classes is consistent; namely, voiceless obstruents are more robust than voiced obstruents (as well as being more frequent members of minimal pairs in general); fricatives are more robust than plosives, but plosives comprise a larger number of contrasts (by a 4:3 ratio); coronals are the most frequent place of articulation and are also the most resistant to perturbation by background noise (21%), while labials are the next most frequent class but by comparison are relatively vulnerable to noise perturbation (24%). Thus, the effect of noise perturbation does not re-rank feature classes in terms of their overall role in minimal-pair contrasts in the lexicon.

Turning to functional load the interpretation of the lower panel of Figure 9 is slightly different from the minimal pair count. Here we can consider higher functional loads to represent the relative importance of the acoustic information underlying a given contrast, because higher ranked phones according to this measure occur in contrasts that are both highly frequent and easily confusable. However, one could also interpret the difference between the baseline and weighted functional loads as representing the relative reliability of a given phone in perception, where [s] and [z], for example, exhibit high phonological functional loads—i.e., they occur in many contrasts in the lexicon whose merger would result in a considerable loss in information—but when weighted by the error rate on such contrasts their functional load drops notably. This means that the sibilants [s, z] play a greater role in the obstruent system in English relative to [t, d, k, h, b], all of which show a higher potential for information loss when perceptual constraints are considered (i.e., their functional load estimates at +2 and –2 dB are higher than are those for [s, z]; the [s, z] FL values drop at both SNRs because while they distinguish many words in English, they are much less likely to be misheard, and thus there is less pressure on their detection in the signal because they are quite robust acoustically).

Overall, according to both measures the plosives and the fricatives [f] and [s] play an outsized role, both at a phonological level, and in terms of how decreases in their probability of detection due to background noise affect the broader discriminative potential of the lexicon. The relative ranking of components of this set does vary depending on the measurement used; however, both measures exhibit the same general partitioning of the system into the core obstruents noted above, and a more marginal set composed of voiced fricatives, affricates, dentals/glottals, and the alveolar flap. Thus, in the next section, where perturbation of specific acoustic cues is simulated, we expect the results to reflect this distinction, prioritizing place and voicing cues among plosives, and manner and sibilance cues between [f, s] and [p, t, k, b, d].

### 5.2 Cue neutralization

Finally, we link the acoustic and perceptual characteristics of obstruent contrasts by studying the impact of cue perturbation on the system of wordform distinctions in the lexicon. The following procedure was adopted for the simulation of acoustic cue perturbation: for each cue in the test set (the *Lex95* database), the contrast parameter was set to 0 (e.g., $\Delta\text{FREQ}_{\text{PK}} = 0$), thus simulating a complete loss of information along that acoustic dimension. Mergers in target parameter values, however, were not simulated, as such a procedure has a tendency rather to capture broad correlations in listener accuracy, such as the better recognition, on average, of voiceless obstruents relative to voiced obstruents, or fricatives relative to plosives. The lexicon model from Section 4 was then used to predict listener accuracy on the cue-perturbed data, with changes in minimal pair count and functional load then used to evaluate the relative role of each cue in the lexicon. Further, this procedure was used as an initial demonstration of how we might simulate the impact of gradient cue loss/weakening on the structure of the wider lexical system, which we hope will be a starting point for future work predicting the trajectory of historical sound changes that are in progress, as well as reconstructing the mechanism behind sound changes that have already taken place.

Figure 10 shows changes in minimal pair count and functional load from individual cue perturbation. Regarding minimal pair counts, the results conform with our initial expectations based on the critical role of plosive contrasts and contrasts with [s] and [f] in the lexicon, as noise duration and F1 provide voicing cues in this set, while consonantal spectral tilt and F2 are robust cues to place of articulation, and relative F3 amplitude cues sibilance distinctions. The functional load results largely agree with this pattern, with the notable exception of preceding vowel duration, a result which derives from the fact that the more restricted set of word-final contrasts are nevertheless highly dependent on

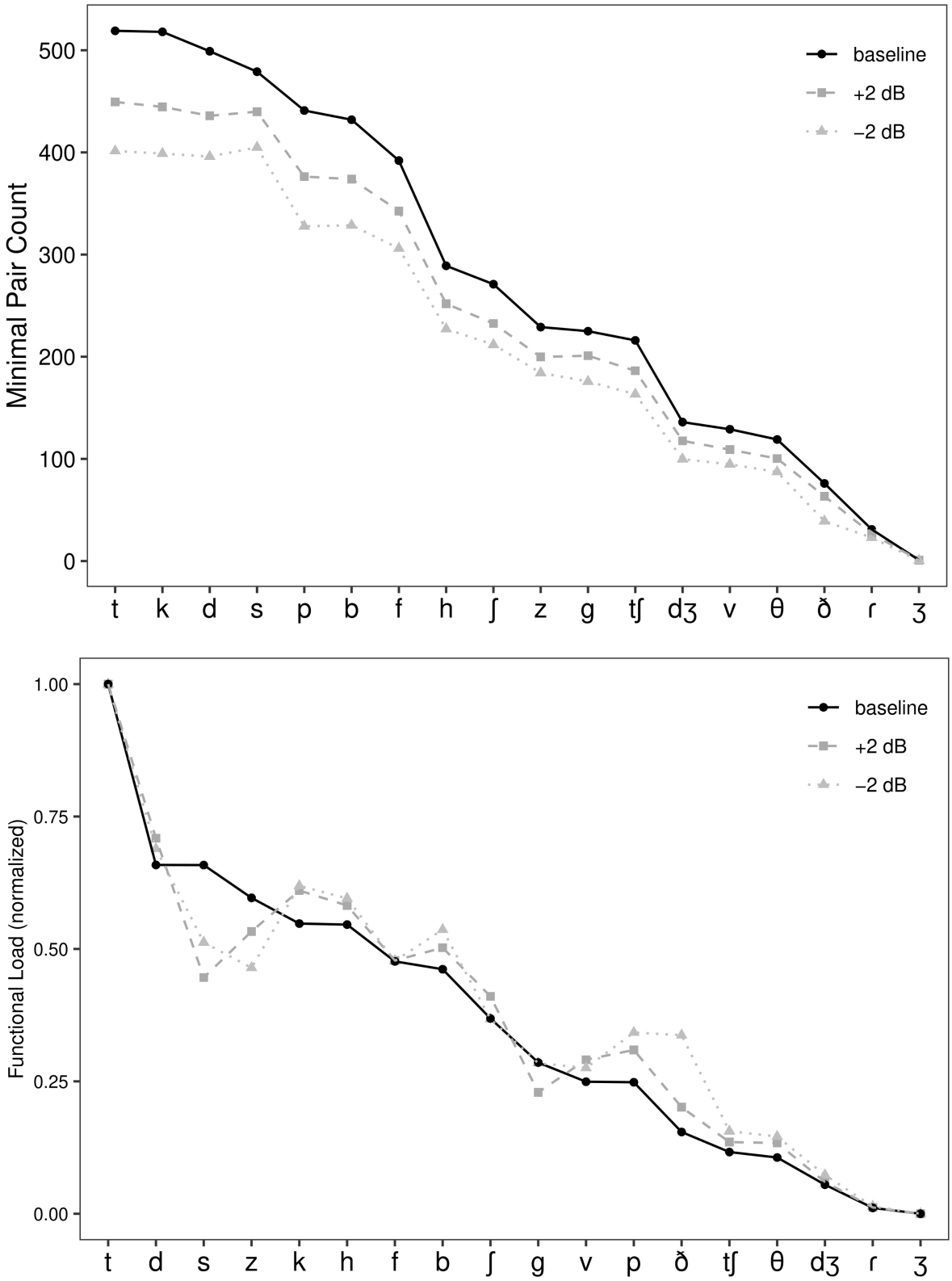


Figure 9: Consequences of auditory perturbation (at +2 and –2 dB SNR) for the number of minimal pairs and functional load of contrasts involving each obstruent phone. Note that phones are ordered according to decreasing baseline values of minimal pair count and functional load, and so the $x$-axis, though similar, is not identical between the two panels.

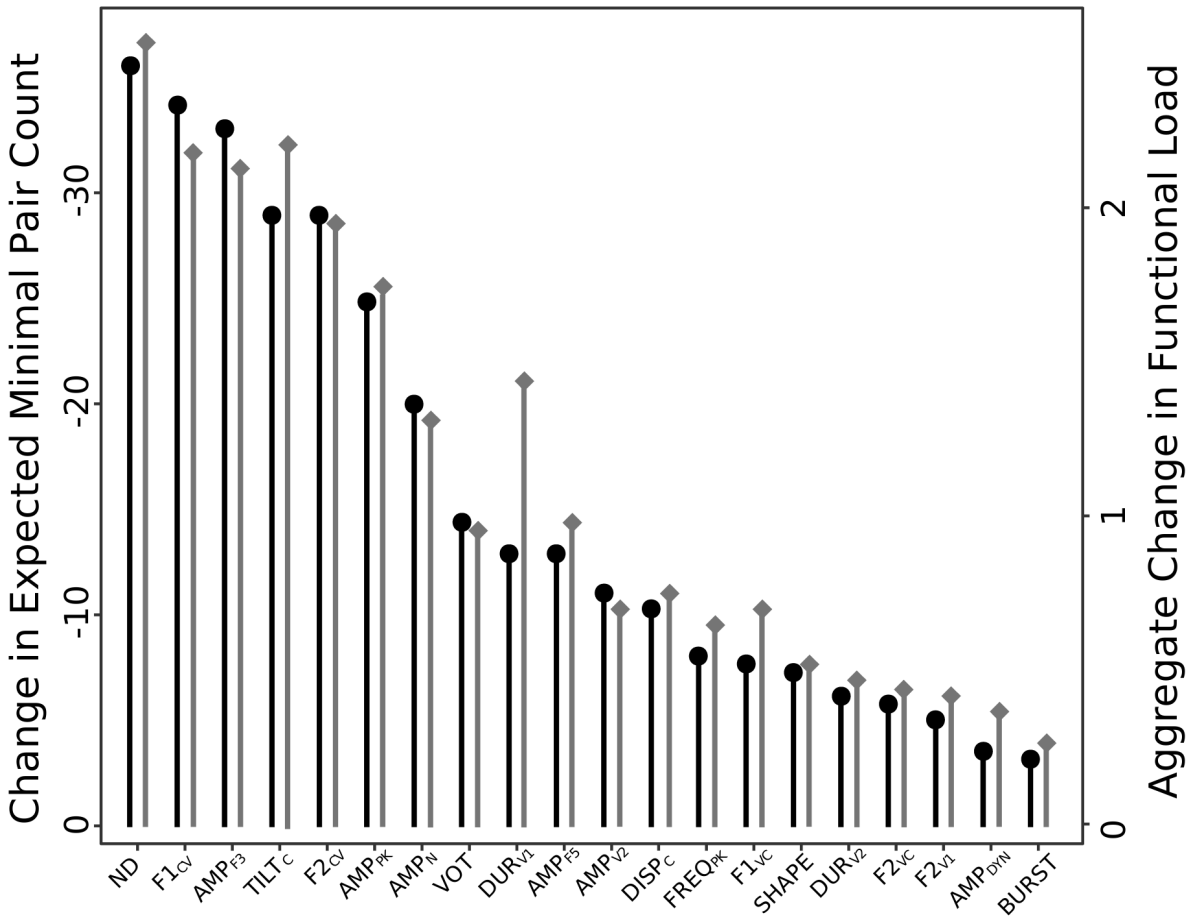


Figure 10: Change in aggregate minimal pair count (black circles) and functional load (gray diamonds) due to perturbation of the twenty most informative cues in the *Lex95* model lexicon. For both cases, higher bars indicate a greater role for that cue in the lexicon, as they reflect greater change in minimal pair counts and functional load.

$DUR_{V1}$ cues, allowing for a notable change in functional load despite a more modest change in minimal pair counts (this is a consequence of the logarithmic form of the FL measure).

In simulating the response of the system of lexical contrasts to perturbation by background noise and cue loss, we were able to identify the critical components of the system from a lexical perspective. That is, given that we know that the distribution of contrasts in the lexicon is highly asymmetric, with the majority of the system dependent on approximately one-third of the obstruent inventory, we asked how such asymmetries affect the impact of acoustic uncertainty on the maintenance of wordform distinctions in English. The result of both the auditory- and cue-perturbation analyses is that the perception of the set [p, t, k, b, d, f, s] by-and-large determines the overall performance of the system in terms of successful message transmission. This means that in measuring the encoding potential of the obstruent system in English, both at a theoretical level and a practical level (such as in the assessment of the impact of hearing impairment on communication), much greater attention must be given to plosive place/voicing contrasts, sibilance distinctions, and manner contrasts within this set. This outcome is not only relevant for synchronic research, but implies similar asymmetries in the diachronic evolution of the obstruent system.

## 6 Summary and conclusions

In studying the phonetic system as embedded in the lexicon, we have adopted a position related to Ohala's (1990) advocacy for integration over interfaces, though our focus has been not on questions of categoricity, gradience, and phonetic grounding, but rather the shared link between phonetics, phonology and the lexicon. By focusing on the

lexicon we introduce a new dimension of inquiry that has gone overlooked in the debate over interface structures and assumptions. Ohala and others have rightly pointed out that there exist many asymmetries in speech that are not captured in all phonological models, but these asymmetries are often addressed at the level of perceptual encoding (Lahiri and Reetz 2002, 2010, for instance, use *underspecification* to account for perceptual asymmetries like the discrimination of coronal versus non-coronal sounds). What remains to be implemented is a framework for handling lexical distributional asymmetries in phonetics, and one that is not derived or secondary as it is in models of speech production and perception (e.g., a model may implement a gradient mapping of acoustics onto lexical entries, but this is considered a downstream implementation of phonetics in lexical access; it does not then feed back onto the structure of the underlying phonetic system). This chapter represents an initial attempt at motivating such a framework.

In the Introduction, we outlined the following argument structure motivating the development of a lexical framework for the analysis of the acoustic and perceptual structure of phonetic systems:

*P1.* The primary basis of phonetic analysis is phonemic.

*P2.* Phonemes by definition perform a lexically discriminative function.

*P3.* The distribution of phonemic contrasts in the lexicon is non-uniform, with contrasts differing in their functional load.

⇒ If the speech system is optimized for transmission of phonologically encoded messages, then perceptual weighting of acoustic information in the signal must reflect this distribution.

The three premises are theoretical deductions from the literature. Their implication, however, is conditional on the assumption of cue optimization for message transmission. While we cannot provide direct evidence for optimization of this kind (to do so would require the comparison of cue integration in multiple linguistic systems), Sections 3 and 4 do provide ample evidence of cue weighting patterns in word recognition that are poorly estimated when distributional asymmetries in the lexicon are not accounted for, and the perturbation analyses in Section 5 provide a picture of the lexicon that is not at all phonetically balanced, and which makes frequent use of a small set of phones, contrasts, and acoustic properties.

This phonetic heterogeneity of the lexicon was foreseen in earlier phonological and computational work, and so the question was not: will similar results be found acoustically and perceptually. The question was what this fundamental fact about the utilization of speech sounds in English word formation would imply for phonetic research programs operating independent of such higher-order distributions. What we find is that there is a wide array of information that is either missing from or distorted in independent, balanced inventory models of the phonetic system. The outsized role of the set [p, t, k, b, d, s, f], and the acoustic properties underlying their discrimination, is masked in a set of 18 phones given equal weight in acoustic and perceptual experimentation. The predominance of multi-feature contrasts (>75%) in the lexicon, which promotes the utility of cues such as F1 over VOT in cross-manner voicing perception, is unanticipated by a model where the balance in phone/contrast distributions makes attending to single-feature distinctions more efficient. And biases toward fricatives, sibilants, and voiceless stimuli are only evident when both the distributional and acoustic characteristics of such sounds in a broad set of real-word contrasts are analyzed. Therefore, the phonetic system must be considered not as an independent component interfacing with other components to various degrees, but as a system embedded within the higher-order systems it encodes.

## Acknowledgements

This work was supported in part by by a Doctoral Dissertation Improvement Grant to the first author from the National Science Foundation (#BCS1918404). We would like to thank Bob McMurray for his insightful comments on an earlier version of this work.

## References

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database. *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA*.

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

Beres, A. M. (2017). Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research. *Applied Psychophysiology and Biofeedback*, 42(4):247–255.

Bermúdez-Otero, R. (2003). The acquisition of phonological opacity. In *Variation within Optimality Theory: Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 25–36.

Bermúdez-Otero, R. (2007). Diachronic phonology. In *The Cambridge Handbook of Phonology*, pages 497–517. Cambridge University Press.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, 2(3):153–164.

Browman, C. P. and Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. *Papers in laboratory phonology I: Between the grammar and physics of speech*, pages 341–376.

Brysbaert, M., New, B., and Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44(4):991–997.

Bukmaier, V. and Harrington, J. (2016). The articulatory and acoustic characteristics of polish sibilants and their consequences for diachronic change. *Journal of the International Phonetic Association*, 46(3):311–329.

Byrd, D. (1995). C-centers revisited. *Phonetica*, 52(4):285–306.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Chomsky, N. (1981). Knowledge of language: Its elements and origins. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 295(1077):223–234.

Chomsky, N. (2017). Two notions of modularity. *On concepts, modules, and language: Cognitive science at its core*, pages 25–40.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row, New York.

Clements, G. N. and Hume, E. V. (1995). The internal organization of speech sounds. In Goldsmith, J., editor, *The handbook of phonological theory*, pages 245–306. Oxford: Blackwell.

Cohn, A. C. (1993). Nasalisation in English: Phonology or phonetics. *Phonology*, 10(1):43–81.

Cohn, A. C. and Huffman, M. K. (2014). Interface between phonology and phonetics. *Oxford Bibliographies in Linguistics*.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences*, 3(3):115–120.

Cooke, M. and Scharenborg, O. (2008). The interspeech 2008 consonant challenge. *Interspeech, 2008*.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

de Saussure, F. (1916). *Course in General Linguistics [Trans. Roy Harris]*. Duckworth, London.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27(4):769–773.

Denes, P. (1955). Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America*, 27(4):761–764.

Docherty, G. J. (1992). *The timing of voicing in British English obstruents*, volume 9 of *Netherlands Phonetic Archives*. Walter de Gruyter.

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, 22(2):109–122.

Dow, M. (2020). A phonetic-phonological study of vowel height and nasal coarticulation in French. *Journal of French Language Studies*, 30(3):239–274.

Evers, V., Reetz, H., and Lahiri, A. (1998). Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics*, 26(4):345–370.

Fant, G. (1960). *Acoustic theory of speech production: With calculations based on X-Ray studies of Russian articulations*. Mouton & Co., The Hague.

Fant, G. (1973). Stops in CV-syllables. *STL-QPSR*, 10(4):110–139.

Farmer, A. K. (1985). Modular theories of grammar. *Annual Review of Anthropology*, pages 25–48.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194.

Fischer-Jørgensen, E. (1954). Acoustic analysis of stop consonants. *Le Maître Phonétique*, 32:42–59.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.*, 84(1):115–123.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4):1357–1392.

Gerstman, L. J. (1957). *Perceptual dimensions for the friction portions of certain speech sounds*. PhD thesis, New York University.

Gray, A. and Markel, J. (1974). A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(3):207–217.

Haggard, M., Ambler, S., and Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America*, 47(2B):613–617.

Hale, M. (2000). Marshallese phonology, the phonetics-phonology interface and historical linguistics. *The Linguistic Review*, 17:241–257.

Halle, M. (1990). An approach to morphology. In *North East Linguistics Society*, volume 20, page 12.

Halle, M. and Mohanan, K. P. (1985). Segmental phonology of modern English. *Linguistic Inquiry*, 16(1):57–116.

Hedrick, M. S. and Ohde, R. N. (1993). Effect of relative amplitude of frication on perception of place of articulation. *J. Acoust. Soc. Am.*, 94(4):2005–2026.

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 33(5):589–596.

Hockett, C. F. (1967). The quantification of functional load. *Word*, 23(1-3):300–320.

House, A. S. and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, 25(1):105–113.

Hughes, G. W. and Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 28(2):303–310.

Jakobson, R., Fant, C. G., and Halle, M. (1952). *Preliminaries to Speech Analysis: The distinctive features and their correlates*. The MIT Press, Cambridge, MA.

Johnson, K. (1997a). The auditory/perceptual basis for speech segmentation. *Working papers in linguistics, Ohio State University Department of Linguistics*, pages 101–113.

Johnson, K. (1997b). Speech perception without speaker normalization: An exemplar model. In Johnson, K. and Mullennix, J. W., editors, *Talker variability in Speech Processing*, pages 145–166. Academic Press, San Diego.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.*, 108(3):1252–1263.

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, 1(6):571–591.

Kenstowicz, M. and Kisseberth, C. (2014). *Generative phonology*. Academic Press.

Kingston, J. (2007). The phonetics-phonology interface. In de Lacy, P., editor, *The Cambridge Handbook of Phonology*, pages 401–434. Cambridge University Press.

Kingston, J. C. and Beckman, M. E. (1990). *Papers in laboratory phonology: Volume 1, Between the grammar and physics of speech*, volume 1. Cambridge University Press.

Kiparsky, P. (1982). Lexical morphology and phonology. *Linguistics in the morning calm: Selected papers from SICOL-1981*, pages 3–91.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7(3):279–312.

Krámský, J. (1974). *The phoneme; introduction to the history and theories of a concept*. W. Fink.

Krull, D. (1989). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *PERILUS*, 10:87–108.

Lahiri, A., Gewirth, L., and Blumstein, S. E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America*, 76(2):391–404.

Lahiri, A. and Reetz, H. (2002). Underspecified recognition. *Laboratory Phonology*, 7:637–675.

Lahiri, A. and Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, 38(1):44–59.

Lehiste, I. and Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *The Journal of the Acoustical Society of America*, 31(4):428–435.

Liberman, A. M., Delattre, P., and Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, pages 497–516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1–13.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 33(1):42–49.

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422.

Lisker, L. and Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, 10(1):1–28.

Lloyd, H. (1936). Note on vowel length. *American Speech*, 11(2):188–189.

Martinet, A. (1952). Function, structure, and sound change. *Word*, 8(1):1–32.

McCarthy, J. J. and Prince, A. (1995). Faithfulness and reduplicative identity. *Linguistics Department Faculty Publication Series*, page 10.

McMurray, B. and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219–246.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27(2):338–352.

Mohanan, K. P. (1982). *Lexical phonology*. PhD thesis, Massachusetts Institute of Technology.

Mugdan, J. (1985). The origins of the phoneme: Farewell to a myth. *Historiographia Linguistica*, 38(1):85–110.

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (2004). TADA: An enhanced, portable task dynamics model in MATLAB. *Journal of the Acoustical Society of America*, 115(5):2430–2430.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America*, 132(6):3980–3989.

Oh, Y. M., Coupé, C., Marsico, E., and Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, 53:153–176.

Ohala, J. J. (1974). Phonetic explanation in phonology. *Proceedings of the Parasession on Natural Phonology, Chicago Linguistic Society*, pages 251–274.

Ohala, J. J. (1990). There is no interface between phonology and phonetics: A personal view. *Journal of Phonetics*, 18(2):153–171.

Ohde, R. N. and Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, 74(3):706–714.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 39(1):151–168.

Peterson, G. E. and Lehiste, I. (1960). Duration of syllable nuclei in english. *The Journal of the Acoustical Society of America*, 32(6):693–703.

Pike, K. L. (1972). *Phonetics: A Critical Analysis of Phonetic Theory and a Technic for the Practical Description of Sounds*. ERIC.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release)[www. buckeyecorpus. osu. edu] columbus, oh: Department of psychology. *Ohio State University (Distributor)*.

Ramchand, G. and Reiss, C. (2007). *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

Redmon, C. H. (2020). *Lexical acoustics: Linking phonetic systems to the higher-order units they encode*. PhD thesis, University of Kansas.

Reetz, H. and Jongman, A. (2020). *Phonetics: Transcription, production, acoustics, and perception*. John Wiley & Sons, 2nd edition.

Rubach, J. (1997). Extrasyllabic consonants in Polish: Derivational Optimality Theory. In Roca, I., editor, *Derivations and constraints in phonology*, pages 551–581. Clarendon.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382.

Scobbie, J. M. (2007). Interface and overlap in phonetics and phonology. In Ramchand, G. and Reiss, C., editors, *The Oxford Handbook of Linguistic Interfaces*, pages 17–52.

Shadle, C. H. and Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. In *ICSLP 96, Proceedings*, volume 3, pages 1521–1524. IEEE.

Shattuck-Hufnagel, S. and Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, 18(1):41–55.

Smits, R., ten Bosch, L., and Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation. *The Journal of the Acoustical Society of America*, 100(6):3865–3881.

Stevens, K. (1992). Lexical access from features. *Speech Communication Group Working Papers*, 8:119–144.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data.

Stevens, K. N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds. *Phonetic linguistics: Essays in honor of Peter Ladefoged*, pages 243–255.

Stevens, K. N. (1998). *Acoustic phonetics*. MIT press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111(4):1872–1891.

Stevens, K. N. and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64(5):1358–1368.

Stevens, K. N., Manuel, S. Y., Shattuck-Hufnagel, S., and Liu, S. (1992). Implementation of a model for lexical access based on features. In *ICSLP*, number October, pages 499–502.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, 3(1):32–49.

Surendran, D. and Niyogi, P. (2003). Measuring the usefulness (functional load) of phonological contrasts. *Technical Report TR-2003-12*.

Surendran, D. and Niyogi, P. (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 279:43–58.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Vandenhoek and Ruprecht, Göttingen.

Tucker, B. V., Brenner, D., Danielson, K. D., Kelley, M. C., Nenadić, F., and Sims, M. (2019). The Massive Auditory Lexical Decision database: Toward reliable, generalizable speech research. *Behavioral Research Methods*, pages 1–18.

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America*, 61(3):846–858.

Walsh, M., Möbius, B., Wade, T., and Schütze, H. (2010). Multilevel exemplar theory. *Cognitive science*, 34(4):537–582.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, 23.

Woods, D. L., Yund, E. W., and Herron, T. (2010). Measuring consonant identification in nonsense syllables, words, and sentences. *Journal of Rehabilitation Research & Development*, 47(3):243–60.