PLASTICITY OF PHONOLOGICAL CATEGORIES

Mark VanDam

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Linguistics
Indiana University
June 2007

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Robert F. Port, PhD

Stuart Davis, PhD

Kenneth J. de Jong, PhD

Tracy Alan Hall, PhD

David B. Pisoni, PhD

June 25, 2007

Acknowledgements

This research, including all that preceded it and what promises to follow, was conducted with invaluable help from so many people and institutions. From inception to completion, this work would not have been possible without the intellectual guidance and patient support of my dissertation director, Bob Port. Generous contributions to my graduate training and dissertation were freely given by the other members of this dissertation committee, Stuart Davis, Ken deJong, Tracy Hall, and David Pisoni. The rich intellectual environment of Indiana University also allowed me to benefit from exceptional teaching, training, and advice from a wide array of unparalleled intellectuals including Diane Kewley-Port, Rex Sprouse, Dan Dinnsen, Paul Newman, Steve Franks, Manuel Diaz-Campos, and Robert Bonte.

Financial support was partially provided by the NIH (T32-DC00012) in the form of a predoctoral trainee fellowship under the direction of David Pisoni and Bob Port.

In no small part, another cohort of folks made important contributions to both the intellectual challenges and social reprieve so critical to completing such a degree: Brian Jose, Tamara Lindner, Jennifer Deegan, Joel Petersen, Matt Start, Noah Silbert, Joshua Herring, Louis Rabaut, Andrea Smidt, Tristan Purvis, Mike Brady, Laurent Bergez, Rebecca Ronquest, Brice Boembeke, Alexa Romberg, Lance Reeder, Brooke Bagwell, Mike Marlo, Susie Levi, Adam Buchwald, Tessa Bent, Rebecca Petrush, Patrick Randolph, Trevor Pittman, and especially Tricia Frank. Many thanks to these people who contributed to my intellectual development and sanity during my training.

Finally, and perhaps most importantly, my family offered patient support at every stage of the process. My parents, Jim and Susan, and siblings, Kip, Marissa, and Kaitlin, were unfailingly supportive and always gave me encouragement.

Mark VanDam
PLASTICITY OF PHONOLOGICAL CATEGORIES

The traditional model for language assumes a level of competence that represents words composed of distinctive features that are invariant units drawn from a small, universally specified set. The present work tests two of these assumptions, representational invariance and the assumed small set size of these features.

This study trained naive listeners on artificially lengthened voice-onset times in specific target words and tested whether pretraining and posttraining perceptual boundary locations differed. Results indicate subjects were successfully trained on the target words, but this effect did not generalize to other, structurally similar words. In addition, robust frequency and lexical effects are reported, but do not apparently interact with training. As predicted, frequency effects suggest subjects preference for high-frequency words over low-frequency words, but, somewhat surprisingly, lexical effects suggest a preference for non-words over words.

Results undermine foundational assumptions of the traditional model, specifically the assumptions of representational invariance and a small set size. To account for these results, an exemplar memory model is discussed in which no principled restrictions are imposed on memory for language. The exemplar models discussed are able to account for the facts reported here.

Table of Contents

1. Introduction. 1
1.0. Two models of linguistic representation: abstract versus exemplar models 1
1.1.0. Architecture of the traditional abstractionist framework. 2
1.1.1. Where do the features come from? 8
1.1.2. How many features are there? The notion of formal *economy*. 9
1.2. The problems of universal features and economy within the traditional abstractionist framework. 11
1.3. A very large number of contrasts shows that economy cannot be maintained. 13
1.4. What can be done to address the problems of plastic features and economy in the traditional model? 18
1.4.1. Rejecting observations incompatible with static features and economy is not a viable alternative. 19
1.4.2. Modification of the traditional abstractionist model renders the model powerless. 25
1.4.3. Replacing the traditional theory entirely is a viable option. 26
1.5. Sketching a rich memory account for language. 27
2. Methods. 30
2.0. Methods overview. 30
2.1. Methods: subjects. 30
2.2. Methods: design 31
2.3. Methods: materials 32
2.3.1. Testing stimulus construction. 34
2.3.2. Training stimulus fabrication. 36
2.4. Hardware. 38
2.5. Software. 41
2.6. Methods: procedure 42
2.7. Methods: data design and analysis 45
3. Results. 46
3.0. Results, introduction. 46
3.1. Results, preliminaries. 46
3.2.0. Results: Session, Set, and Group effects on boundaries. 48
3.2.1. Results: lexical frequency 53
4. Discussion. 59
4.0. Overview of discussion. 59
4.0.1. Summary of main results. 59
4.1. Plasticity of linguistic representations. 59
4.1.0. Framing the time scale of linguistic change in speech perception. 60
4.1.1. Verbal transformation modifies category perception at a very short time delay. 63
4.1.2. Selective adaptation modifies category perception. 64
4.1.3. Categories shift in response to shadowing or immediate recall tasks. 70

4.1.4. Categorical shifts are observed after exposure to higher-order linguistic stimuli. 74
4.1.5. Categories shift as a result of accommodation within a discourse. 76
4.1.6. Categories shift as a result of exposure in the perceptual learning paradigm.79
4.1.7. Summary of categorical plasticity 83
4.2. Categorical plasticity effects considered in the present study. 84
4.2.1. Category plasticity is observed after training. 84
4.2.2. Rate of category shift is fast: only a small number of exposures is required for category shifts. 86
4.2.3. Successfully trained subjects do *not* generalize training to new lexical items. 87
4.2.4. Importance of training effects. 91
4.3. Frequency effects of the present work considered. 91
4.3.1. There is an asymmetry in the perception of voiced and voiceless frequency.92
4.3.2. Frequency differences in the voiced series. 96
4.3.3. Auditorily ambiguous non-word bias: a possible anti-Ganong effect. 100
4.4. Interaction effects among experimental factors: do frequency effects interact with training effects? 103
4.5. The traditional model versus exemplar model. 105
4.5.1. Plasticity of features can be accounted for in an exemplar model. 106
4.5.2. Linguistic memory including richly specified features can be accounted for in an exemplar model. 107
4.6. Summary of results and implications. 108
5. Appendix 110
6. References 116

List of Figures

Figure 1. The traditional generative language model 6
Figure 2. Physical hardware configuration and subject location. 40
Figure 3. A typical identification function for all data from one subject. 48
Figure 4. Main effects of training. 50
Figure 5. The interaction effects of Group, Set, and Session. 52
Figure 6. VOT boundary values by lexical frequency specification. 54
Figure 7. Voiced and voiceless VOT boundary locations for lengthened training group, training lexical set only. 55
Figure 8. Voiced and voiceless VOT boundary locations for shortened training group, training lexical set only. 56
Figure 9. Boundary VOT values for the group trained on lengthened tokens is shown.. 57
Figure 10. Boundary VOT values for the group trained on shortened tokens is shown. 58
Figure 11. Schematic response curve shift after selective adaptation. 68
Figure 12. Schematic response curve shift after perceptual learning. 82
Figure 13. Effects by group (shortened, lengthened), lexical set (training, generalization), and session (pretraining, posttraining, generalization). 89
Figure 14. Lexical frequency effects of words by voicing. 93
Figure 15. from Miller, Green, and Reeves (1986)... 95
Figure 16. from Ganong (1980).. 98
Figure 17. Lexical effects on VOT boundary for voiced word.................................... 101

List of Tables

Table 1. Alveolar target words used in pretest, training, and posttest phases. 33
Table 2. Velar target words used in pretest, training, and posttest phases. 33
Table 3. Alveolar target words used in the generalization phase. 33
Table 4. Velar target words used in the generalization phase... 33
Table 5. Sample phrase segments from training story 37
Table 6. The organization of word pairs by lexical frequency. 92
Table 7. Alveolar training production sentences. 110
Table 8. Velar training production sentences. 111
Table 9. Supplemental training production sentences. 112
Table 10. Alveolar generalization production sentences. 113
Table 11. Velar generalization production sentences. 114

1. Introduction.

1.0. Two models of linguistic representation: abstract versus exemplar models.

There are two models of linguistic representation that will be discussed in this dissertation: the traditional abstract model and an exemplar model. The traditional language model was described at least as early as the Port-Royal Grammar (Lancelot & Arnauld, 1660). This model became the foundation of 20th-century linguistics as codified by Saussure (1916, 1996), Sapir (1921), Jesperson (1924), Bloomfield (1933), Harris (1951, 1962), Chomsky (1957, 1965, 1966), and many others. These foundational works led directly to a detailed phonological model typified by Jakobson, Fant, & Halle (1952/1963) and Chomsky & Halle (1968) [henceforth, *JFH* and *SPE*, respectively]. *JFH* and *SPE* were especially important foundational works contributing to the development of the traditional model because they explicitly detail the assumptions and architecture of the model, and, in particular, the critical role of *abstraction* in the traditional phonological model. Hence, the traditional model is fairly uncontroversially an *abstractionist* model. For purposes of the present work, I will refer to the *traditional* and *abstractionist* models interchangeably.

The consequences of the abstractionist account are twofold for the present work. First, several specific properties of the abstractionist assumptions are directly tested by the experiments reported here. The experimental results are in turn considered against the assumptions of a different model, an *exemplar* or *rich memory* model. Since those models make clear and contrasting predictions about the outcome of the experiments reported here, the results will bear on the theoretical viability of each model. Second, in a more general sense, the abstractionist model has had a deep impact on the study of psychology of speech production and speech perception. Although the abstractionist model has been shown to be incompatible with a large body of empirical results, a subset of which will be described below, the details of the abstractionist model are important to consider head-on to clearly identify how the present work bears on the assumptions of that model.

This chapter briefly describes the abstractionist model with attention to a subset of critical assumptions, giving special attention to the notions of linguistic *economy* and representations of *universal features*[1] in memory. A number of abstractionist criticisms are offered, as well as possible approaches for dealing with those criticisms. A description of an alternative approach, a rich-memory or exemplar model, is described to account for data that are incompatible with the abstractionist model. Importantly, the contrasting theoretical accounts of how language is stored in memory make different predictions about how people use language. The present work is intended to directly test which of these predictions accounts for the empirical data of a controlled set of experiments specifically designed to test those assumptions.

1.1.0. Architecture of the traditional abstractionist framework.

The traditional language model has three parts, the lexicon, the grammar, and the performance. The lexicon is a static cache of information specifying the words of the language as the arbitrarily paired meaning and form of each lexical item (Lancelot & Arnauld, 1660; Pierce, 1867; Saussure, 1916). The specification of the form includes syntactic specifications, the sequence of phonemes, and so on. At the level of sound

[1] The notion *universal features* properly contains the idea that these features are invariant. The specific criticism described here is directed toward invariance, but since the universality of features is a broader term, it will be used here.

representation, a lexical item is hypothesized to consist exclusively of segments defined by their distinctive features (DFs). DFs are idealized, abstract, discrete, atomic, static symbols that are arranged in serially ordered, segmented vectors as the atoms of sound structure.[2, 3] The segmented DFs are the exclusive representational units for lexical storage. The grammar[4] is a collection of obligatory instructions or rules that transform lexical forms. The grammar is absolutely naive to the meanings associated with lexical items. That is, the phonological grammar performs transformations exclusively, fully, perfectly on the forms that are passed to it without respect to either the lexical meaning or the performance. Critically, the units used by the grammar are the same DFs used for lexical representation. Together, the lexicon and phonology comprise *competence*. In competence, the lexicon "should contain only idiosyncratic properties of items, properties not predictable by general rule,...[consisting of] just enough information for the rules of English phonology to determine its phonetic form in each context; since the variation is fully determined, the lexical entry must contain no indication of the effect of context on phonetic form" (*SPE*, p. 12). Stating that only 'just enough information' is specified is the result of the assumption that memory, including linguistic memory, is limited. That is, the storage capacity of humans for language is assumed to be very small, and thus must

---

[2] Here I am explicitly referring to DFs as the basic, atomic units ('formatives' in *SPE*) used in competence for sound structure. DFs, however, are clumsy to deal with because they are abstractions, in most cases unpronounceable or impossible in isolation. For example, the DF [±coronal] is only able to be formally identifiable as a unit, but is neither formally nor physically independent, having a variety of DF co-occurrence restrictions and requirements. To avoid confusion and simplify descriptions, DFs are arranged in ordered columns (a 'feature vector', or, with multiple segments in serial succession, a 'feature matrix') "using letters of the alphabet as informal abbreviations for certain complexes of features, i.e., certain columns of a matrix...taken to be a two-dimensional matrix in which the columns stand for consecutive units and the rows stand for different features" (*SPE*, pp. 10, 294).

[3] Here I am assuming a representation that does not formally address the issue of feature dependence or the relatedness of features. There have been attempts, such as Feature Geometry, which have attempted to address this issue.

[4] The grammar properly contains a number of subparts such as the syntax, morphology, and phonology. The present discussion will be primarily restricted to the phonology.

be minimally specified. This is accomplished by positing abstract features from a very small set of primitives. Thus, the basic propositional elements of the model are DFs arranged into serially ordered segments which are the exclusive units for representation in the lexicon and transformational operations in the phonology.

An important property of so-called *universal features* is that they are invariant over time, between and within language users, and formally identical when recurring within the lexicon. Language users might use different features, or representations might consist of different features, but, it is claimed, the features themselves are universal and do not change. Universal features must be absolutely static in the abstractionist model and "represent the phonetic capabilities of man and, we would assume, are therefore the same for all languages" (*SPE*, p 295). Notice that this is a claim about the minimal units of representation, but not the amalgamation of features into larger units such as morphemes or words. The morphemes and words of course vary between languages (and users), but the atomic units of representation, the features, are hypothesized to be absolutely identical. The surface implausibility of two different languages using absolutely identical basic sounds is accounted for by the competence-performance distinction described in detail below. Essentially, observed differences are explained away by suggesting that those differences are only performance differences while the units of phonological competence among languages are identical. The theoretical assumption of universal features critically bears on the claimed strengths of the traditional model in areas such as language acquisition, typological and cross-language comparisons, and a potential biological basis for language (Hauser, Chomsky, & Fitch, 2002, Fitch, Hauser, & Chomsky, 2005). Despite the potential explanatory benefits of the claim of universal features, however, the basic assumption may not be itself empirically motivated.

Given the basic units and their assumed universal nature, how are they used for language? Upon selecting a lexical item (normally as the result of talker intent), the item is passed from the lexicon into the grammar and manipulated according to grammatical rules in domains including syntax, morphology, and phonology. The symbolic string is said to be totally determined within linguistic competence, and, while within competence, is represented only by the abstract set of units available in that domain, the DFs. Once the string is determined, it is passed out of competence and into the performance module which articulates the string in physical terms, which, under normal conditions, is always speech. The model is based on a production model, but perception can be accomplished only by the physical (mostly auditory, but also visual) system recognizing or recovering the abstract feature units from sound. The assumption of universal units among talkers and listeners is an advantage to the model here, since recognition is the recovery of feature units produced from identical code, never only similar units or even slightly different units. Perception essentially reverses the flow of information to recover the shared universal units of competence and ultimately the meaning associated with the lexical item (or talker's intent). In this system, language *talis qualis* is exclusively restricted to the domain of competence, and performance is not in the domain of language but is governed by processes formally irrelevant to linguistics. Figure 1 gives a visual display of this architecture.

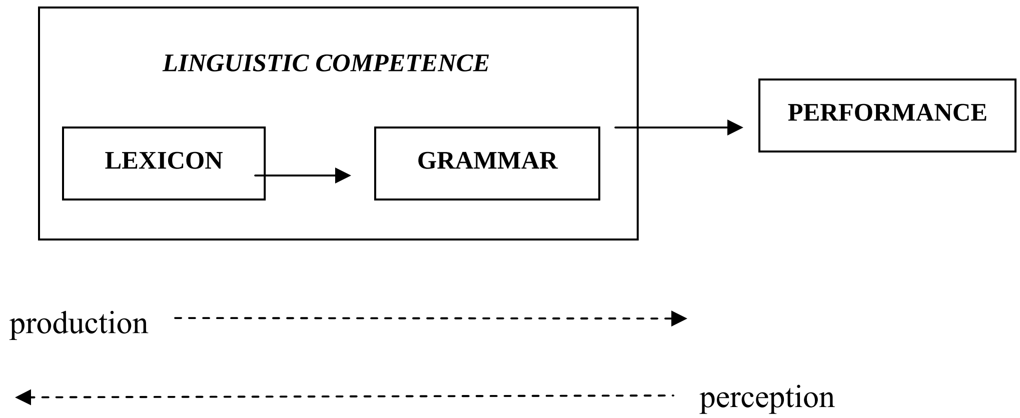

Figure 1. The traditional generative language model

The problems dealt with in phonetics are often regarded by non-phonetician linguists as motor descriptions, physical events orthogonally or incidentally related to speech, automatic implementation rules, or simply outside of linguistic competence and thus not formally part of linguistics or even language study. Under this traditional view, the phonetic system is not critically related to the abstract, universal units of language in competence, but phonetics is a (necessary) performance, perceptual, or articulation component which interfaces with the symbolic units of competence only inexactly. The grammar performs the necessary transformations according to the syntax, morphology, and phonology, each higher level passing its results to the next lower level, until the output of competence is passed along to be articulated by the physical performance system. The performance domain is formally naive to the competence domain, handling only what is passed to it without feedback or knowledge of its meaningful content or transformations carried out.

Competence, however, cannot be directly observed but must be inferred from performance since that is the only point of access (Chomsky, 1965; Pisoni, 1997). This apparent paradox—performance is simultaneously required for language but defined to be outside language—is a matter of long-standing interest and debate in the literature. Working in the abstractionist tradition assumes that there must be an interface to mediate between competence and performance. But after many years of research, there is no consensus on the exact nature of that interface, or even whether competence and performance can in principle inform each other. As described in greater detail below, the failure to resolve this issue is in part due to the failure to question the primary assumptions and subsequent reliance on strict abstraction as a propositional starting point.

Despite the abstractionist paradox and subsequent failure to rectify the problem, the abstractionist account has not totally ignored certain aspects of performance. Phonological research, believed to be about the competence using the undisputed feature units, typically assumes the units (DFs, in most cases) with formal properties and leaves it up to the phonetician or other researcher to describe the exact nature of the universal apriori units, including their physical description. If an invariant physical description in the performance domain eludes precise correspondence with the presumed feature, it is not typically interpreted as evidence against the feature (by some phoneticians as well as many phonologists), but rather as evidence of experimental inadequacies, or to noise in the performance domain. No matter where the failure originates to find invariant definitions for all the phonetic features, it cannot be contained in competence, because competence is axiomatically assumed to possess only those universal features.

1.1.1. Where do the features come from?

Within competence, the traditional model posits a set of indivisible, atomic DFs as arranged in segment-sized vectors. A specific DF is argued to be justified when a contrast persisting into the lexicon is demonstrated. This demonstration is normally accomplished by presenting a minimal word-pair such as *time* and *dime* which differ by a single DF (the DF 'voicing' in this case) while all other features are identical. If a contrast cannot be accounted for by physical or deterministic means, the source of the contrast is posited as a DF and admitted to competence, available to the phonology, and, more generally, joins the set of universal feature primitives. It was thought that the discovery of DFs in a variety of languages would converge on similar contrasts that would ultimately reveal a single, comprehensive set of universal DFs that would define the phonetic capabilities of humans (Port & Leary, 2005). Although no language would likely use all the available DFs to represent words in the lexicon, only DFs from the universal set would be possible as units in the representational lexicon or the phonological grammar. The universal set of DFs used by every language must be a formal subset of the universal set, and must be fixed to a small number of units so that distinctions between DFs can be guaranteed. The lexicon must be formally perfect (in computation, recognition, transformations, etc.), and it is clear that formal perfection can only be achieved if the units are sufficiently distinct from one another. Thus, the abstractionist account posits a small universal set of DFs that are the only possible units in competence. Lexical representations are compared exclusively from the DFs employed in that language, and the phonological grammar is entirely restricted to manipulations and transformations on those DFs (although, to account for allophony, the grammar might require a slightly larger set of DFs than the lexicon—but all DFs must come from the universal set).

1.1.2. How many features are there? The notion of formal *economy*.

Given the theoretical hypothesis that there is a universal set of DFs from which all languages draw a subset, what is the set like, and, specifically, how many features are there? Although this account has no theoretical way of directly restricting the number of contrasts that will actually be observed or the total number of DFs, it *assumes* that a small number will be found. This notion of a restricted DF base is known as *economy*. But assuming economy does not justify it, so a number of ad hoc reasons including descriptive simplicity, computational tractability, typological universality, and elegance have been offered. More important than these ad hoc reasons, however, is the fact that without economy, the abstractionist theory becomes less general and less powerful.

A phonological rule in the grammar is powerful and productive if it applies to many lexical items. A rule is weak and contributes little to overall competence if it applies to a very small number of lexical items. If lexical representations are allowed to be more richly specified due to an increase in the number of features, more phonological rules and more complex rules will be required to apply to a greater variety of lexical forms. This problem is addressed in *SPE* (somewhat incompletely) with the admission that "a huge number of ad hoc rules" (p. 295) is undesirable and references to a reduction in 'value' or an increase in 'cost' when the number of rules approaches the number of lexical representations (*SPE*, Chapter 7). For example, suppose two lexical items share a feature in the more economical model. The application of a single phonological rule targeting the shared feature affects both lexical items. If, however, the same lexical items no longer share that feature but are represented by two different (perhaps perceptually or acoustically similar) features, then two rules would be required where only one was necessary in an economical model. There is in fact a quantifiable limit on the relationship between the number of features and phonological rules: as the number of features approaches the number of rules, the rules increasingly fail to add explanatory power to the model and can be dispensed with. The power of a rule is contained in its potential to apply to any lexical item that meets the rule's structural description. If a rule applies to a structure that a lexical item does not share with any other lexical item, then the phenomenon can be adequately captured in the lexicon without any rule at all. In short, the abstractionist model for language requires a relatively small number of abstract units for it to have reasonable theoretical power.

Although an exact number of features has never been offered for the competence module, the *JFH* model contains a total of only about 12 DFs for all consonants and vowels and the *SPE* model contains about 40 DFs for all languages, so a number on the order less than one-hundred seems a reasonable maximum.[5] If, on the other hand, thousands or tens of thousands of DFs are required, then economy is clearly violated and the problem of theoretical adequacy discussed above must be considered. It would not likely be problematic for the traditional abstractionist theory if a small number of DFs

---

[5] Although they do not have such a rigorous definition of features (nor, perhaps, a clear understanding of a separation between sounds and orthographic letters), the Port-Royal grammarians make a strikingly similar statement apparently addressing economy. They describe "the marvelous invention of composing out of 25 or 30 sounds that infinite variety of words" as a key element of their grammar (Lancelot and Arnauld, 1660: 22).

had to be added infrequently as new observations warranted. However, it would be a serious problem for the traditional abstractionist account if new observations routinely require large numbers of new DFs.

An important effect of the assumption of economy is that linguists looked for contrasts describable with a small set of DFs, but largely ignore phonetic differences outside traditionally proposed possible contrasts. Linguists chose not to look at phonetics in detail, but assumed a fixed, small, economical set of contrasts as a point of departure. Economy further constrained the sorts of directions that the abstractionist theory could look to in order to account for new observations not immediately compatible with the basic feature inventory. In the abstractionist account, linguistic variability has to be constrained to the predefined feature units, while variation not describable with those units is, according to the model, ignored or disallowed within the language system. Naturally, variations of all kinds are recognized outside the linguistic system in performance. For example, although the traditional model recognizes that a stop may vary in a large number of ways, it may only vary in a very small number of ways linguistically, in terms of the DF units available to the phonology.

1.2. The problems of universal features and economy within the traditional abstractionist framework.

The notion of universal features is potentially problematic for the traditional abstractionist model. Simply, the theory strongly predicts that the universal features used for representation and in the phonology are unchangeable and absolutely fixed. Testing this basic assumption is simple and straightforward; for example, if features and contrasts exhibit dynamic, flexible, or plastic representational structure, fixed universal features cannot be reasonably maintained. Equally problematic for the traditional model would be violations of linguistic economy, which will be described in detail below.

The assumptions requiring economy in traditional abstractionist phonology can also be tested. If a large number of contrasts are necessary, the theory must admit each one as a feature of some sort. But for every admission, economy is violated, and the power and usefulness of abstractionist theory is reduced. In the following section, several classes of examples are offered that require just this sort of increase in the number of features required. Each example presented below describes a contrast requiring lexical specification. Lexical contrasts, as opposed to a contrast that can be derived by phonological rule, are important because the ways in which contrasts are shown depend on unpredictable features of a word. Recall that the phonology proposed by the abstractionist model contains only the rules for performing transformations, but no lexical representations itself. The lexicon, on the other hand, contains representations of totally unpredictable forms.[6] If a contrast is totally predictable, then that contrast can be accounted for in the phonology exclusively, but if a contrast is shown to be unpredictable or word-specific, it must be represented in the lexicon. Also recall that the abstractionist account must posit shared units in the competence generally, so whether a feature is argued to originate in the lexicon or in the phonology affects competence equally.

---

[6] The lexicon, in this model, is likely constrained by factors not considered here. For example, given totally unpredictable forms of serially ordered segments, we might expect a form such as /tktk/ to be as likely as /tɑtɑ/. The former is, so far as I know, not attested in any language while the latter is widely attested. The notion of unpredictability, however, is intended to suggest the arbitrary association between form and meaning (Saussure, 1916). For example, there is no reason whatever that we associate the form /mɑɹk/ with the image of the author; the image might just as well be represented by another linguistic form, such as /fɹɛd/ or /mæɹi/ or /ɹʌmpəlstɪltskɨn/.

Examples problematic for the traditional approach are described below, including non-Neogrammarian sound change, lexical frequency, and sociological factors.

1.3. A very large number of contrasts shows that economy cannot be maintained.

Beginning in the late 19th-century and throughout the 20th-century, linguists—especially phoneticians, sociolinguists, and speech perception psychologists—observed many cases where various language facts can not adequately be described using the traditional DFs such as those later used in *JFH* and *SPE*. This precipitated the kind of problems for the assumption of economy in the traditional abstractionist model described immediately above. Perhaps the first objections came from those challenging the Neogrammarians in the late 19th century. Neogrammarians suggested that sound changes were the result of systemic changes in the phonological grammar that would naturally affect any sound occurring in the lexicon that met a particular structural description. It was assumed that the lexicon and phonological grammars were distinct from one another and should not be expected to influence each other (recall that the phonological grammar is totally automatic or 'formally naive' to the contents of the lexicon; that is, as far as the phonology is concerned, it is passed a string of features and it applies its transformations to those features without regard to *any*thing else). Thus, any sound change (as a change in a rule in the phonological grammar) should rather discretely affect all lexical items to which it applies without respect to any property of a lexical item other than its DF specification. Some linguists pointed out that sound changes diffuse more gradually over the lexicon affecting certain words before other words despite both words meeting the same structural description (Schuchardt, 1885; Wang, 1969; Labov, 1972, 1981; Hooper, 1976; Phillips, 1984; Bybee, 2001, 2006). If some words are affected before others and the specification is not otherwise predictable from distinctive phonetic features already expressed in the lexicon, lexical specifications must be expanded to specify which lexical items are subject to the phonological rule and which are not. Since a previously unadmitted DF must be added, the principle of economy is violated and theoretical power is reduced.

Another notable example of the necessity of a large number of DFs is the familiar lexical frequency effect (Schuchardt, 1885; Howes, 1954, 1957; Savin, 1963; Hooper, 1976). The frequency of occurrence of a lexical item has been shown, in a host of linguistic contexts and in a variety of linguistic domains, to affect language production and perception. Some lexical items are used more frequently than others, and are pronounced and perceived in empirically different ways from one another. For example, Hooper (1976) observes the English words *memory* and *mammary* which share almost all structural features, but, due presumably to usage frequency differences, the former may undergo syncopation (loss of a medial vowel) while the latter may not. Similarly, Ganong (1980) showed that listeners were more likely to identify an acoustically ambiguous token as an actual lexical item as opposed to a non-word when presented with a minimal-pair contrast such as *peef-beef* or *peace-beace*. Critically, when presented with a minimal-pair asymmetry in which one form is an actual lexical item and the other is not, listeners show a perceptual bias toward the actually occurring lexical item. These data suggest a bias to recognize real words over non-words. The work of Hooper and Ganong is representative of a very large body of literature showing robust usage effects.

To account for usage phenomena, the abstractionist model must specify the source of the differences in competence, presumably in the lexicon, by creating new abstract universal features since those universal features are the only way it is allowed to represent such variation. The only way to obtain unpredictable phonological differences in the traditional model is to have those differences specified in the lexicon. In the syncope described by Hooper and lexical-bias described by Ganong, but also applicable more generally to a variety of similar usage-based results, the traditional account is obligated to posit a DF to specify each observed unpredictable variation. And in each case, the principle of economy is violated and the overall theoretical power of an abstractionist is reduced.

The last group of examples come from a wide-ranging and vast literature on phonetic change in language reported in the sociological and sociolinguistic literature. The description and examples below depend on taking seriously the necessary assumptions of representation in the traditional model. By taking the assumptions seriously, the traditional model can be directly assessed and criticized. In the traditional model, a lexical item is a paired meaning and form. Different lexical meanings are associated with different forms, and any form is a universal-feature matrix of serially ordered segments. A feature is added to the inventory when it is thought to be the source of difference in forms that have different meaning, such as the feature [VOICE] in the minimal pair *pug* and *bug*. Of course, a minimal pair is one straightforward kind of evidence when the two words are different in a precise, specific way so that the source of difference is isolable to a single segment. Notice that the requirement for difference (in meaning or form) must be simply that that difference is detectible by actual language users—but never to the degree of difference or degree of contrast since the features, by the demands of the model, must be perfectly discrete. Insofar as the form difference is merely an arbitrary label for the meaning difference it signifies, the only requirement here is that language users do in fact produce different forms that mean different things. Thus, the discussion below, describing results from the sociolinguistic literature, will show differences in meaning that are detectible to language users. By the rules set out in the traditional model, then, these points of contrast must be admitted to the universal distinctive feature set, necessarily undermining the assumed economy and universality of their distinctive feature set.

At least as early as Mencken (1919), who described subtle differences between American and British English varieties, explanations (or descriptions) of sound changes requiring a very large number of phonetic distinctions have been offered. Critical sources of contrast have been argued for the social status of talkers (Labov, 1963, 1966), varying levels of politeness (Goffman, 1955; Brown & Gilman, 1960), sex and gender of talker (Bloomfield, 1927; R. Lakoff, 1973), race (Wolfram, 1969; Labov, 1969; Dillard, 1972; Burling, 1973), and host of other factors. If differences such as these sociological factors (such as race or gender—and there are surely many hundreds more described in the literature) are represented in linguistic memory, they must be admitted into the universal feature set when they are shown to be distinctive. The level or degree of distinction is formally irrelevant to the traditional model since it is by definition only able to deal in discrete, abstract, feature-sized units for all representation. For example, as shown by Labov's famous study of New York City speech in department stores (1966), talkers controlled their productions of rhoticization based on perceived social factors of the interlocutors. A department store employee produced a word such as *four* in different ways based on (the perception of) the talker. Since these talkers produced different meanings and different forms—albeit slightly different meaning and slightly different forms—those differences must be represented in the abstractionist model in the only possible way it may account for differences, with features.

The fact that there is a consistent difference, for any reason, is the prima facie justification for it to exist. Actually, it is more than prima facie evidence, it is an absolute requirement of the theoretical model. As described in detail below, the theory cannot choose features to include or not to include or evaluate the content of the feature. Determination of features is totally deterministic, never teleological. That is, a feature is said to be required when, for example, a minimal pair can be offered as evidence. The phonological feature is responsible for representing that lexical difference, but neither the phonology nor the lexicon is allowed to have access to the contents of the other. Furthermore, a large number of robust and regular sociological effects like these described above is well attested in the literature. As such, each demonstrated sociological effect, each reliable distinction, must be admitted into the feature set of linguistic competence, again undermining economy and reducing overall theoretical power.

It deserves notice that these examples above are best understood as broad ranging, widely applicable, and extendable not only across but within languages. That is, language use details such as lexical frequency and social status of talker seem unlikely to be restricted examples, but much more likely to be found in all communities and among all talkers. Supporting this claim, these sorts of robust effects have been found in virtually every investigation when researchers have looked closely at the details of a language. As mentioned above, the thriving fields of phonetics and sociolinguistics are a strong testament to the necessity of looking at features that, as a collected whole, provide repeated evidence that violates the notion of economy crucial to the traditional model. Similarly, they can be taken as evidence that humans detect and store much of these differences.

1.4. What can be done to address the problems of plastic features and economy in the traditional model?

The proposition of a fixed universal feature set described above is a critical assumption of the traditional model. The examples given above—non-Neogrammarian sound change, lexical frequency, and sociological factors—are evidence in favor of memory for language that uses a large number of distinctions, which must be understood as a large number of DFs in the abstractionist theory. But too large a number of DFs undermines the viability of the traditional abstractionist theory because it violates the principle of economy. So, given these serious challenges to the traditional abstractionist model, the traditional linguist must choose one of several options: (a) a rejectionist approach which ignores, dismisses, or discounts challenging observations, (b) a modification approach which modifies the abstract, traditional account to be compatible with the observations, (c) a replacement approach which disposes of the fundamental assumptions of the traditional abstractionist account. As will be argued below, the most promising option is the replacement approach, replacing the traditional assumptions with a theoretical perspective whose architecture is well-suited to account for facts consistent with richly specified memory. Each of these options will be considered below.

1.4.1. Rejecting observations incompatible with static features and economy is not a viable alternative.

The first approach a traditional linguist might make in response to the challenges described is to reject the challenges themselves. There are essentially three types of rejectionist arguments. In one approach, the challenges are not recognized at all. If a researcher does not recognize the objections, no coherent stance can be assembled. Since this is essentially burying the head in the sand, there is little to say about it. Second, a linguist might reject the challenges to the traditional abstractionist approach head on, on theoretical grounds. The general idea in this case would be to accept the architecture and assumptions of the traditional model, with a strictly demarcated, narrow competence component separate from the performance component. In this view, items of linguistic competence remain the universal, abstract units subject to the linguistic rules, and everything else is formally undefined with respect to language. If an observation (of any kind) is undefined in terms of the formal linguistic system, that system has absolutely nothing to say about it in any regard. This is essentially an attempt to deal with incompatible observations by defining them out of the problem (Newmeyer, 2003, 2006; Bromberger & Halle, 2001; Hale & Reiss, 2001). This approach, too, is of minimal interest since any potentially problematic observation is dispatched by simply defining it into a module that lies outside of competence, thus not relevant to linguistics. This view is perhaps the most common approach taken by linguists committed to the traditional perspective. In general, this view is not sympathetic to making the DF inventory larger or the units flexible but rather assumes the inventory is fixed and essentially a closed class with all necessary units axiomatically stipulated.

The third sort of rejectionism is more complicated and nuanced, since its proponents suggest that the problems described above are not really problematic since serious, practicing linguists do not take the formal, abstractionist assumptions seriously, and because of this the results and conclusions of linguistics do not critically depend on the assumptions described above. This perspective essentially accepts the rejection of the traditional assumptions, but further contends that it is not worthwhile to continue to consider the traditional model at all. They contend that the characterization of the traditional view given here—especially citation of *JFH* and *SPE*—is an anachronistic straw-man perhaps appropriate several decades ago but not now. These responses, however, cannot be seriously maintained by even a cursory investigation of the literature and current trends in linguistics, especially phonology, but inside phonetics research, too. There are repeated, clear examples of linguists explicitly or tacitly accepting the assumptions of universal features and economy criticized above. To demonstrate this claim, I will offer here a few examples of recently published work that exhibit what Scobbie (2005) calls "the broad church of generative grammar" including Stevens (2002), the rise and prominence of Optimality Theory, and Hale and Reiss (2001).

Stevens' paper (2002) serves as an updated and expanded exposition of his quantal theory (Stevens, 1972, 1989), and, to a considerable degree, is an attempt to reaffirm the phonetic theory of *SPE*. Stevens' hypothesis is that certain regions of acoustic and articulatory space are more stable than other regions. These regions of stability are where discrete, invariant speech sounds are located, and, importantly, these 'quantal regions' of stability correlate with a small number of invariant, discrete distinctive features that in many cases resemble those posited in *SPE*. In this theory, speech communication is successful if talkers and listeners are able to perform within the quantal regions where continuous variation (or sloppiness) is mapped on to a small number of universal categories or features. Stevens' quanta discretize linguistic events in a way consistent with a model that demands, among other things, economy of features.[7]

Another illustration of the importance of the traditional abstractionist perspective is the rise and present status of Optimality Theory (OT) in phonology. OT and the traditional view are both abstractionist theories of language and share the notions of universal features and economy outlined above. Both OT and the traditional rule-based models assume DFs as the representational units for lexical memory, so the differences in the two approaches lie in the phonological grammar component of competence. The rule-based model assumes the basic tenets of formal propositional logic (syllogism, discreteness, etc.) while OT assumes a different method of transforming the DFs passed from the lexical representation which allows for, among other things, violations of formal constraints (or rules). The differences in specific grammatical transformation methods are not important for the present discussion since both the traditional model and OT assume the same formal properties including economy and a fixed universal feature set,

[7] Quantal theory predicts that regions of stability will be found, but is not necessarily formally invested in maintaining economy. However, even though the theory does not require economy, Stevens seems to be interested in interpreting quantal theory in terms explicitly compatible with *SPE*-style language modeling. That is, it appears that although quantal theory does not strictly require economy, quantal theory does not undermine economy, which is in turn interpreted as support for *SPE*-style models that do critically depend on economy.

among many other properties.[8] However, the assumptions of propositional logic directly borrowed into abstractionist models have received less critical attention than the violable constraint system of OT, so OT researchers have given somewhat more attention both to the nature of the proposed constraints and the evaluation processes in the grammar itself. For example, John McCarthy, one of the most influential proponents of OT, describes the need for limiting the number of constraints: "Positing a new constraint is not to be undertaken lightly. Constraints in OT...are claims about UG with rich typological consequences.... [Positing] new constraints may seem to offer an easier solution, but they can bring a cost in typology," (2002: 39).[9] Since constraints are an indispensable part of OT grammar, McCarthy seems to be warning against bringing too many things into competence and reducing the economy of the representations. McCarthy is observing that proposing a new constraint is a claim that that constraint is admitted to competence and should be expected to manifest itself in various typological descriptions. Thus, McCarthy takes very seriously indeed the notions of fixed universal features and economy that the abstractionist model demands.

As a barometer of the field, I examined the contents of two top journals in theoretical linguistics published in 2006, *Phonology* and *Linguistic Inquiry*. Of the 24 total research articles in those two journals, 15 articles (62.5%) made reference to OT, and most of these critically rely on OT for the analysis including tableaux, ranked

[8] Other shared formal properties include perfect transformational execution, timeless and rational computation, discreteness of tokens, real-time memory and processing restrictions, and so on. For an extended discussion of language modeling as a formal system, see Port & Leary (2005).

[9] The use of *typology* here is meant to refer to language types as the totality of ranked combinations of constraints (the phonological grammar in OT). Each proposed constraint is hypothesized to be an element of Universal Grammar (UG), and must therefore be ranked with respect to every language. Admitting a new constraint requires that new constraint to be ranked in each language, and thus, based on different rankings among languages, the addition of a constraint predicts different phonological grammars, or different languages types.

constraints, and the like. Notice, too, that in the case of *Linguistic Inquiry*, the focus of the research articles is not exclusively in the domain of phonology where OT is strongest and its application could be argued to be influenced by current research trends and topics considered fashionable. Since OT demands abstractionist assumptions of the universality of non-language-specific constraints and economy, the heavy use of OT offers undoubted (but usually tacit) endorsement of the tenets required by the traditional abstractionist language model.

In another recent work, Hale and Reiss (2001) espouse a complete separation between phonology (competence) and phonetics (performance) as described above. To them, performance is mechanistic, automatic, and by definition not part of competence. This work is an example of taking the critical assumptions of the abstractionist hypothesis very seriously indeed. Taking their admittedly 'extreme formalist position,' they make repeated reference to what they see as the scientific need for Occam's Razor, Fodorian modularity, elegance, and 'mathematical explicitness' as critical factors in determining or assessing the things that they see as linguistic competence versus the things they see as extralinguistic performance and thus unable in principle to inform linguistic issues. They "assume that the phonetic substance (say, the spectral properties of sound waves) that leads to the construction of phonological entities (say, feature matrices) *never* reflects how the phonological entities are treated by the computation system....Phonology is not and should not be grounded in phonetics" (p. 167 [emphasis in the original]), and assert that "phonology should be all form and no substance" (p. 173). In the sense they are talking about phonological grammar here, 'all form' means precisely that they define phonology as formal (i.e., rational, in the Cartesian sense) and abstract, without any physical or articulated necessity of substance whatsoever. They suggest that since some aspects of speech are not linguistic, by extension, no aspects of speech can be linguistic. One primary reason they make this assumption is to satisfy the notion of economy and limit the number of things in the linguistic competence. It is critical that economy be respected so that the competence is able to compute the linguistic outputs which it can pass to the performance (although they are totally unconcerned with performance). Likewise, they tacitly accept the notion of universal features, or their model would not be a formal model at all. At some point it is assumed that some representation is passed to the performance mechanism, but that performance is formally unaffiliated with the competence. Overall, Hale and Reiss seem to be essentially advocating a dead-end for all experimental language research, or even empirical observations of any kind about language. They disconnect what they claim is interesting—linguistic competence—from all directly observable phenomena. This is another example of modern theory and thought in the discipline of linguistics taking the idea the abstract traditional model and concomitant universal features and economy very seriously.

This review of the state of theoretical linguistics shows the pervasive influence of the traditional model on modern linguistics, and gives evidence of a variety of research lines that argue against the relevance of any observable data in order to maintain the larger theoretical perspectives. In short, this approach allows the theory to trump empirical observations, and even, in the case of Hale and Reiss (2001), simply defines any problematic data as non-linguistic by fiat. The difficulty, then, is how one can verify or refute a theoretical model when all evidential data have been rejected as irrelevant.

1.4.2. Modification of the traditional abstractionist model renders the model powerless.

Another approach for trying to salvage the traditional model is to attempt to retain the critical bulk of the traditional model and its abstractionist underpinnings but modify it (as little as possible) so that it accounts for apparently incompatible observations. The present discussion is particularly interested in how the assumptions of universal features and economy pose problems for an abstract linguistic model. As described above, both assumptions are central to the most basic mechanisms underlying an abstractionist model. The modification approach asks what the consequences are of removing those assumptions from an abstract model. The answer is simply that an abstract model completely falls apart, losing all of its power.[10]

If the traditional abstractionist model really worked, the outcome would be spectacular: given a very small input with fewer than 100 or so DFs, all of the important portions of every output form would be totally determined. If this were true, once the phonology were determined, the work of phonologists would be done with nothing more to say. But if universal features and economy constraints are removed or relaxed, the number of phonological generalizations required increases with every new DF. The number of generalizations must necessarily exceed the number of DFs, so new DFs, as the only available competence units for which the rules may act upon, would require a grammar (i.e., the totality of ordered rules) incomprehensibly large. Since the

---

[10] The notion of *power* refers generally to descriptive power, explanatory adequacy, and ability to predict what forms might and might not occur. In general, a theory or model is said to be powerful if it is able to achieve broad predictability and generalization through a small set of observations. For example, if a single observation applies to all forms, that observation is powerful. A theory or model is said to have less power if it applies to fewer forms, captures fewer generalizations, and/or is able to predict fewer outcomes or forms. For example, if a single observation applies to a single form, that observation has little power.

competence model assumes critical limits on computational power and memory, the large grammar would constitute an intractable problem (Idsardi, 2006). Overall, modification of an abstract model is unreasonable not because modification is unreasonable per se but because in this case the content of the modifications required would reduce the abstract model to powerlessness. Or, put another way, the assumptions of a fixed, universal feature set and economy are so central to the abstractionist traditional model that their removal would decimate the theory and reduce it to nothing.

1.4.3. Replacing the traditional theory entirely is a viable option.

Finally, if the traditional theory encounters problems that it cannot adequately deal with, a final option is to supplant that theory with another theory. To some degree it is a matter of interpretation how seriously to consider data that appear inconsistent with a theory, and dispensing with a theory too quickly risks throwing the baby out with the bathwater.

In this case, however, the criticisms above invite serious consideration of the two primary issues of universal features and economy. Those assumptions must be more thoroughly investigated as theoretical assumptions. If universal features and economy are shown to be incompatible with the facts of language, the viability and usefulness of the traditional abstractionist theory is itself in question. Granting doubt about the viability of the traditional model as described above, an alternative model must be proposed. In this case, a theoretical alternative might assume a perspective directly addressing the assumptions criticized above. For this purpose, a rich memory account is described below in which essentially no memory constraints on the lexicon or the phonological grammar are supposed, permitting a potentially vast number of flexible features for speech.

1.5. Sketching a rich memory account for language.

So far it has been argued that observed facts are incompatible with the traditional assumptions of universal features and economy. Of course, impassioned rejection of *SPE* phonetics is not a new perspective (Lisker & Abramson, 1971; Sampson, 1974), and others have offered more broad criticism in the field of linguistics (G. Lakoff, 1973; Ross, 1973; Labov, 1972; Johnson, 1990, 1997; Port & Leary, 2005). If the observed facts demand that we dispense with specific assumptions and the concomitant traditional abstractionist account, we must adopt a replacement account that not only deals with the issues previously problematic, but also accounts for a respectably large portion of the observations the traditional abstractionist model deals well with. That is, replacement of one theory with another is risky if the new theory addresses a smaller number of criticisms than problems it creates.

To account for lexical diffusion, the Ganong effect, lexical frequency, and a host of other factors traditional abstractionist models fail to adequately account for, recent work has suggested rich-memory models[11] to account for language production, speech recognition, and linguistic memory (Goldinger, 1996, 1997, 1998; Goldinger, Pisoni, & Logan, 1991; Johnson, 1990, 1997, 2005, 2007; Pierrehumbert, 2001, 2002, 2003; Hawkins, 2003; Pisoni, 1990, 1997). Although many details need to be worked out, the

[11] For present purposes, the term *rich-memory* is a cover term for a class of models that do not assume abstract DFs or lexical economy as a critical feature. Many models, along with their names, interpretations, and applications, have been proposed for this type of model including *exemplar*, *episodic, high-dimensional,* etc.

essential claim is that speech events are experienced, stored, and accessed in a very richly detailed memory that includes both rich sensory and contextual details as well as many categorical descriptive properties. Not every aspect of experience is necessarily stored, but there is no principled limit to what is stored (in terms of kind of thing, such as a DF, or in terms of amount of thing, such as repeated exposures to similar events in the world). New episodes and linguistic experiences are admitted to memory and stored with a vast number of features as clouds of representations in multi-dimensional memory space. Each stored exemplar has a decay function which determines the strength or accessibility of an exemplar representation over time. Categorization can be accomplished by evaluating similarity based on a relevant cloud of partially overlapping exemplars. Any exemplar likely belongs to many categories and several organizational clouds. When a new exemplar is encountered it is added to memory, and, as explained by Gahl & Yu, "all exemplar-based models assume that each experience alters the entire category system slightly" (2006: 213).[12]

The rich memory approach has certain advantages over the traditional abstractionist approach. In particular, it can account for the criticisms above and the issues addressed in the present study—but it also poses many problems for understanding speech perception and spoken word recognition. The main problem is that the rich memory approach places far greater demands on memory than the economical approach based on abstract phonological representations. Since speech is undeniably complex

[12] There are a number of details here that are distracting to the more general discussion, such as precisely how an exemplar updates memory. As explained by Wedel, "depending on the model, new experiences are assigned to relevant categories by comparison to actual exemplars (Nosofsky, 1986), or to generalizations emerging from the exemplars that make up a category (Hintzmann 1986; Goldinger 1996; reviewed in Tenpenny 1995)" (2006: 252). Additionally, these additions/alterations could be compatible with connectionist models, relevant to activation levels, and a host of other instantiation issues and models.

within and between organizational levels, any theoretical account designed to account well for some phenomena will probably find difficulty to adequately explain at least some other phenomena (Merton, 1936, 1949). However, given the critical state of affairs that abstractionists theories find themselves in and the growing body of incompatible evidence, it is important to look closely at the viability of other approaches such as the rich memory approach to account for this evidence.

Important for the present purposes is that the issues of phonological economy and universal features critical to the abstractionist model are directly testable. The present work is designed to address these issues through a series of experiments that test the notion of universal features and the principle of economy. If lexical representations can be shown to critically depend on rich lexical specifications, that evidence will offer support for the rich memory model and add to the literature of observations problematic for the traditional abstractionist perspective. The present work looks closely at the representation of the voicing contrast as a function of voice onset time. Linguistic representations of the voicing contrast are examined through subjects' perceptual identifications of category membership in response to finely controlled stimuli varying along a VOT continuum. Detailed representational structure is then examined through a number of linguistic variables. Systematic representational variability will support a rich memory model for language.

2. Methods.

2.0. Methods overview.

In order to more closely examine memory for words in language use, subjects performed an initial perception task to determine the perceptual VOT boundary of a specific list of word pairs. They were then exposed to a set of synthetically altered target words over the course of five days, and finally retested on those target words as well as some new words to determine how much the effect of exposure might generalize. The abstractionist theory described above predicts that limited exposure to a small number of words with lengthened or shortened VOT should not influence perceptual boundaries at all. Thus, no effect on target words or on generalization words is expected under this model. Exemplar memory models, however, predict some changes in VOT boundary on the target words but not on the generalization items, since identification judgments rely on concrete representations of recently heard tokens of items to be identified. Linguistic productions by the same talkers were collected in addition to perceptual data. Although the focus of the present work is on speech perception, it is likely that there are task-specific ramifications of the training procedures that indirectly might affect production. Therefore, the production procedures are sketched below, but the data are left to future work for examination.

2.1. Methods: subjects

The 13 female and 8 male subjects were recruited on campus at Indiana University. Male subjects average 31.4 years old, females 25.3 years old. All subjects were monolingual native speakers of general American English born in the United States. Although all subjects had at least some foreign language experience, none reported near-native proficiency or fluency in any language other than English. None reported extensive contact with foreign language speakers or extended experience outside of English speaking communities in the United States. None reported a history of speech or hearing deficits, all had normal or corrected-to-normal vision, and all were right-handed. Subjects were paid for their participation.

The 20 subjects[13] participated in all tasks, randomly assigned to one of two training groups differing only in VOT of target words during the training phase of the experiments. One training group consisted of six females and four males, the other of seven females and three males. All subjects were naive to the purpose of the experiment. Additionally, four trained linguists participated as volunteer judges for stimuli naturalness assessments.

2.2. Methods: design

Although the experiments included a broader design to investigate both production and perception, the present report is primarily concerned with perception. Future reports will include the production data independently and in consideration with the perception results described here. However, since production tasks were undertaken by the same subjects in the same sessions as the perception experiments, there is a possibility that the production tasks might have influenced the perceptual data. All subjects performed identical production tasks.

[13] One recruited subject was excluded from the analysis due to inability to perform consistently. This subject failed to identify category boundaries in the posttest phase of the experiment, in some cases, unable to show asymptotic selection choice for more than half of the target words. Analysis of his data revealed patterns of consecutive 'voiced' or 'voiceless' selections without variation over several tokens, indicating he did not perform the task as instructed.

2.3. Methods: materials

Target words were used in two ways, as categorical end-points on a rhyming-pair continuum such as *time* to *dime*, and embedded into a short story for use during the training portion of the experiment. There are a total of 58 target words used during the experiment, selected based on frequency estimates from the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis 1984) and balanced by place of articulation. The stimulus set includes words of very high frequency such as *town* (greater than 100 occurrences per million), very low frequency such as *gawk* (less than 1 occurrence per million), and non-words or 'no frequency' words such as *dalc* (phonotactically licit non-words). Alveolar and velar place of articulation are equally represented. Finally, after crossing voicing, place, and frequency it was necessary to construct two comparable word sets, one for use during training and the other for posttraining experiments to test for possible effects of generalizability.

Target words are presented in Table 1 through Table 4 below arranged by frequency, voicing, and place of articulation. Words in Table 1 and Table 2 are used in pretesting, during the training, and in the posttesting. Words in Tables 3 and 4 are presented to subjects in the generalization portion of the experiments, only after all other material has been finally presented. A set of ten supplemental words used in the testing phase are described below.

Table 1. Alveolar target words used in pretest, training, and posttest phases.

ALVEOLAR	no-Frq [d]	lo-Frq [d]	hi-Frq [d]
lo-Frq [t]	*talc ~ dalc*	*tine ~ dine*	*teal ~ deal*
hi-Frq [t]	*take ~ dake*	*time ~ dime*	*town ~ down*

Table 2. Velar target words used in pretest, training, and posttest phases.

VELAR	no-Frq [g]	lo-Frq [g]	hi-Frq [g]
lo-Frq [k]	*kiln ~ giln*	*caulk ~ gawk*	*cot ~ got*
hi-Frq [k]	*keep ~ geep*	*call ~ gall*	*could ~ good*

Table 3. Alveolar target words used in the generalization phase.

ALVEOLAR	no-Frq [d]	lo-Frq [d]	hi-Frq [d]
lo-Frq [t]	*teak ~ deak*	*taupe ~ dope*	*ton ~ done*
hi-Frq [t]	*top ~ dop*	*ten ~ den*	*too ~ do*

Table 4. Velar target words used in the generalization phase.

VELAR	no-Frq [g]	lo-Frq [g]	hi-Frq [g]
lo-Frq [k]	*kith ~ gith*	*coo ~ goo*	*curl ~ girl*
hi-Frq [k]	*can ~ gan*	*car ~ gar*	*came ~ game*

2.3.1. Testing stimulus construction.

To create the stimuli for the testing continua, two stimulus talkers produced each pair of rhyming words in naturalistic citation forms in the laboratory. Productions were then edited to construct a VOT continuum from 20 to 76ms in 8ms steps, yielding 8 discrete steps designed to cross the categorical boundary between voiced and voiceless.

One male and one female talker not also serving as subjects volunteered to provide natural citation productions of each word for stimulus preparation. Productions judged as highly natural by an external panel of four trained judges[14] were selected for manipulation. The average natural voiceless VOT was 71.4ms for the female talker and 75.5ms for the male; the VOT of the voiced word was not recorded because it was excised and not used in the experiments. Each target pair was accessed in a waveform/spectrogram editor (Adobe Audition 2.0) and manipulations were made in the following way. The rhyme of the voiced member of the pair, minus 3-5 glottal pulses at the beginning of vowel production, was excised and placed into a new file. The closure, burst, and following period of voicelessness were then copied from the voiceless member of the pair and pasted onto the beginning of the rhyme of the voiced member in the new file. In all cases of waveform manipulation (cutting and pasting) zero-crossings were matched to maintain naturalness. This resulted in a form with the exact VOT value of the original voiceless member and the rhyme of the voiced member minus 3-5 glottal pulses.

The VOT of the constructed form was then measured. If the VOT was greater than 76ms, non-adjacent small portions (about 2ms) of the voicelessness were excised to achieve a VOT value of 76ms. If the VOT was less than 76ms, a small portion of the voicelessness was copied and pasted back into the voiceless portion in non-adjacent

[14] Volunteer judges included one advanced graduate student and three PhD phoneticians.

segments. Adjustments from the natural VOT were required in almost all cases. This resulted in a synthetic form consisting of nearly all of the rhyme from the voiced member of the pair and 76ms of VOT from the voiceless member of the pair. These constructed forms with 76ms VOT were then presented randomly to the four trained judges interleaved with original, unmanipulated voiceless tokens. Judges could not reliably distinguish constructed from natural forms above chance.[15]

To prepare successively shorter members of the continuum from this 76ms VOT standard, 8ms of voicelessness was repeatedly excised from the waveform. In each case, very small non-adjacent portions (1-2ms) were cut from the voiceless portion of the VOT, and, in addition, a very small portion (about 1ms at each step) of the high-energy burst portion was excised. Cutting in this fashion (both from the burst and the period of voicelessness) was done in an effort to maintain maximum naturalness. Each progressively shorter token was created in this fashion from the immediately longer token in the continuum so that adjacent tokens contain a maximum of shared signal information. Each stimulus token was individually stored with 200ms of silence (captured from the original recording) immediately preceding the initial burst and following the last glottal pulse or release of the rhyme. In total, 24 continua with 8 steps

[15] Volunteer judges, all trained phoneticians, were asked to indicate whether a presentation was 'natural' or 'synthetic'. They were told that some tokens were natural and some were synthetic. Each of the 24 lexical items was played four times, twice each for the natural and synthetic stimuli for 96 selections for each of four judges (384 total responses). The presentation order was entirely random (determined by a randomization algorithm generated on the fly in Matlab). Pooled responses from the four judges are displayed in the confusion matrix below. Correct responses (true-natural 0.481 and true-synthetic 0.483) and incorrect responses (false-natural 0.517 and false-synthetic 0.519) are roughly at chance performance (0.5 response values), with mean correct response slightly below chance (arithmetic-geometric mean is 0.482).

		actual	
		synthetic	natural
response	synthetic	98	105
	natural	94	87

each for each of 2 talkers yielded a total of 384 stimuli for use in the pretest, posttest, and generalization test.

2.3.2. Training stimulus fabrication.

For the training stimuli, a short 600 word story (given in full in the appendix) was prepared to contain each of 30 total target words exactly twice. Target words include all words from Table 1 and Table 2 above (except the four non-words, to preserve semantic coherence and intelligibility) and a supplemental set of ten words: *tint, taint, tuque, tot, took, cull, kale, kip, keel, come*. The ten supplemental words are not directly tested in the perceptual report here, but were included to increase the sample size for forthcoming reports on related production experiments.

The same male and female talkers used for the testing continua stimuli recorded careful productions of the training story, *The Selfish Recruit*. Productions were then edited, including splicing portions from different renditions of one production into another, editing noise and extraneous sounds, and normalizing and filtering the signal. This process resulted in a cleaned, highly natural master production of the training story for the male and female stimulus talkers, each about four minutes long. The master productions were then cut into 74 phrase-sized segments, each segment a few words long in which one or two target words occurred, but never occurred either initially or finally. The segmentation scheme for each stimulus talker was identical. Several sample phrases are given below in Table 5. Target words are underlined here, but were not in the version presented to subjects.

segment	phrase
27	*and should not buy talc for everyone*
48	*back over the keel segment of the boat*
70	*and get dye to tint some of the caulk for their project*

Table 5. Sample phrase segments from training story

Similar to stimuli used in the continua, 200ms of silence was added immediately preceding and following the first and last words in each phrasal segment. Natural VOT values for each target were recorded. Exact copies of the segmented story were then manipulated using audio editing software (Adobe Audition 2.0). Lengthening and shortening the target words' VOT was performed similarly to the preparation of the testing continua stimuli (i.e., augmentation or excision was performed in very small increments at strategic portions of the signal), except that the natural rhyme of each target was used because training tokens were constructed only within the voiceless series, obviating the need for voiced tokens or tokens approaching the voicing boundary. This procedure also maintained a very high degree of naturalness since the whole of every target word was produced in its intended, original context. Since context is maintained, factors known to influence speech production and perception such as prosody, intonation, and other temporal factors were strictly limited. To create the lengthened VOT of target words, natural VOT values were lengthened to 180% of natural, and for shortened VOT natural VOT was shortened to 80% of natural. These values are somewhat arbitrarily chosen as benchmark VOT values that could be substantially altered while maintaining the perception of naturalness. That is, it was determined that lengthening by about 180% or shortening to about 80% of original was not immediately perceived as unnatural.

In the training story each target word is immediately preceded by an open syllable and followed by a word beginning with a fricative. This method frames each target word by preceding sonorant and following fricative allowing for ease (and possibly later automation) of measurements. Since each target word begins with a stop—or more precisely, a stop closure—a preceding sonorant is acoustically dissimilar and distinct, so boundary points are relatively easy to determine and consistently identifiable. Similarly, since no target word terminates in a fricative, high energy spectral noise clearly and uniquely identifies the starting point of the following word. This frame served another important purpose. In contrast to words used for the preparation of continua, target words used in the training session were embedded in coherent, running speech. The focus of this work is on VOT which is inherently a variable measure dependent on rate of speech, among other things. By allowing for a standardized measure of total word duration (including word boundaries), we are able to take relative speech rate into account. Although relative speech rate is not critically important for the perception experiments reported here, production recordings were gathered in the experiments, and this methodological consideration will facilitate forthcoming production reports as well as comparisons between perception and production measures.

2.4. Hardware.

All phases of the experiments were conducted in an acoustically treated recording room in the Linguistics Department at Indiana University. For all portions of the experiment, stimuli recorded in a single-channel were duplicated into right and left channels and presented to each ear simultaneously via headphones (Sony MDR-V15). Volume was adjusted to be comfortable for each subject. Subjects held a button box used to advance the audio presentation (described below) with both hands out of contact with the table. Presentation of stimuli were controlled via a dedicated personal computer (AMD Athlon 1800+ processor, 768MB DDR2100 RAM, AOpen Cobra DW 850 Deluxe sound card interface). The line-level sound signal was broadcast via the soundcard directly to the headphones (with an additional split for a second pair of headphones available to the experimenter). Volume level was controlled via software associated with the soundcard.

For the perceptual boundary test portions of the experiment, subjects were seated directly in front of a 17-inch LCD computer monitor (MaxStar A170E1 T5) at a comfortable viewing distance (approximately 50cm) with a keyboard on the table below the monitor. The keyboard was modified so that two same-sized keys on the lower right portion were labeled A and B. Additionally, the usual 'P' and 'X' keys were labeled PAUSE and QUIT (which, although available for those functions, were never used by a subject).

For the training portion of the experiment, subjects were seated at a table with a Shure SM-58 microphone placed on a table stand approximately 30 degrees off-center directed at the subject's mouth. The microphone continuously recorded all utterances onto a DAT recorder (Tascam DA 40) after preamplification (Mackie Microseries 1202-VLZ). The presentation equipment (including PC, button box, and audio playback equipment) was independent of the recording equipment (microphone, preamplifier, recorder). The interface given to subjects was a three-button button-box. The left button, labeled REPLAY, replayed the current utterance of the story, the right button, labeled

ADVANCE, initiated playback of the next segment, and the unlabeled middle button was not used.

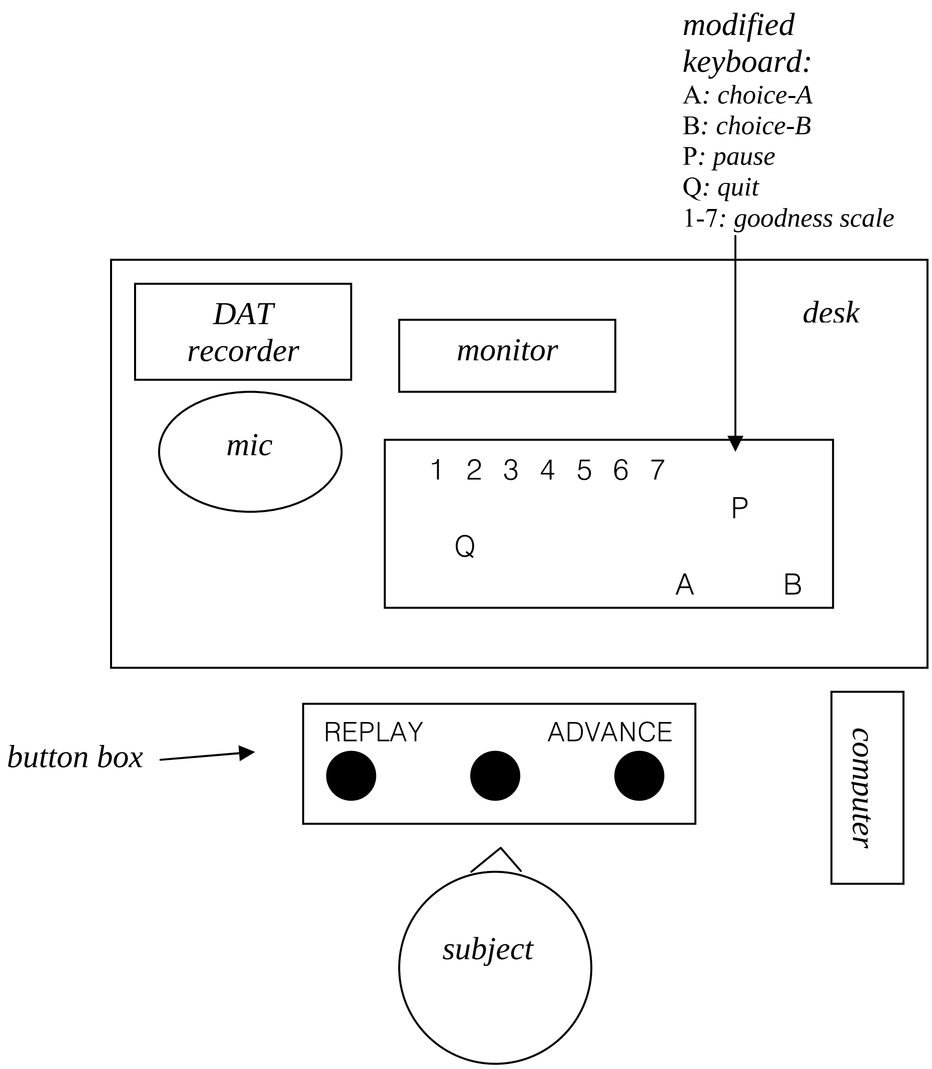

Figure 2. Physical hardware configuration and subject location. The primary display and recording equipment was situated on the table, the computer was on the floor under the table, and the button box (when in use) was operated from the subject's lap.

2.5. Software.

The PC controlled all audio and visual presentations via task-specific software written in Matlab. For the pretest and posttest, two rhyming words were presented at the top of the screen in large black typeface on an off-white field in a two-alternative forced-choice identification task with labels corresponding to the keys on the modified keyboard, for example A for DEAL and B for TEAL. For each pair, the voiced member of the pair corresponded to the A button and the voiceless member to the B button. Presentation order was blocked by pair and randomized on the fly across the two talkers, word pair, and synthetic VOT value within each word pair. Subjects were presented with all eight synthetic VOT values of a single word pair by the same talker, but the order of VOT values, the order of words, and the order of stimulus talkers were randomized. After a subject identified a word, the next audio stimulus file was played and the screen displayed the relevant identification choices. At the bottom of the screen on every trial subjects were given the option to press buttons labeled PAUSE or QUIT. No replay option was available.

For training, each stimulus was played, then a 500ms 350Hz tone. Subjects were instructed to wait until after the tone before repeating the phrase just heard. Pressing the REPLAY button replayed that stimulus and following prompting tone. The replay option could be initiated as many times as requested on the button box. Pressing the ADVANCE button advanced to the next audio presentation in order. Subjects could not return to a presentation after pressing the ADVANCE button.

For all production phases, an orthographic prompt was presented on the screen. Subjects articulated the text on the screen, used the button box to advance, and were continuously recorded by the DAT. Short breaks were offered after 100 productions or five-minute intervals, whichever came first. If a subject selected a break, all data collection equipment was paused.

2.6. Methods: procedure

After introduction to the equipment, procedures, and broad overview of experiment participation, subjects were given a short practice test to illustrate the procedures, after which they were given opportunity to resolve any procedural questions.

Subjects performed pretest production and perception tasks. For the production task, subjects read on-screen presentations for target words in three contexts: (1) isolation, (2) the carrier phrase *say _____ very clearly*, and (3) in a semantically coherent phrase. Presentation order was pseudo-randomized so that the same target was never presented within 4 presentations of itself, and the same context was never repeated more than twice in a row. All pretest phrases are given in Table 7, Table 8, and Table 9 in the appendix below. In all contexts except the isolation condition, target words appeared after an open vocalic or sonorant segment and before a fricative. As described above, this frame is intended for standardization and potential automatic data collection. Each of the 36 target words was produced four times in each context (twice in each of two different coherent phrase contexts), yielding 432 pretest productions from each subject.

For the perceptual task immediately following the production task, subjects were instructed to identify a word played over headphones using the keyboard as quickly and accurately as possible in a two-alternative forced-choice identification test. After the identification they were instructed to indicate confidence on a Likert-style scale (Likert

1932) at the 1-7 number keys (which had modified labels indicating only the relevant number and no other symbols). The procedure was self paced, but after an identification was made the next presentation was automatically initiated. No feedback was provided. The complete pretest lasted about 45 minutes. Pretests were conducted on the first day before training.

After pretraining testing, subjects were randomly divided into two groups for training purposes. One group was trained on the story with synthetically lengthened VOT and the other on the story containing synthetically shortened VOT target words. Except for the manipulated VOT of the target words, the training procedures and materials were identical for both groups of subjects, except for some randomization of presentation order.

For the training procedure, subjects were instructed to listen to an audio-only presentation of each of the 74 phrases, wait for a prompting tone, and repeat the utterance in a normal speaking voice as naturally as possible. The 74 phrasal segments were presented in the same order as the story twice for each stimulus talker on each day. A subject heard a phrase segment such as *and should not buy talc for everyone*, a short 500ms, 350Hz tone, and then simply repeated the presentation. The order of male and female stimulus talkers was randomized by day (either male-female-male-female or female-male-female-male). After each production subjects pressed a button labeled ADVANCE on the button-box to advance to the next presentation. Repetition utterances were continuously recorded to DAT tape.

Subjects were encouraged to repeat an utterance fragment for any reason (e.g., cough during repetition production), but were not given any specific instructions regarding performance, attention, or criteria for controlling the stimuli. Subjects were also instructed to replay the current utterance as many times as desired. It was initially thought that subjects would use the replay function regularly, but replay was rarely used, especially after the first trial through the story. The procedure was self paced, and a presentation was initiated immediately upon pressing any button. No feedback was given. On each of five training days subjects heard and repeated *The Selfish Recruit* four times, twice each from each of two stimulus talkers for a total of 20 listen-and-repeat cycles, hearing each lengthened or shortened target word and then producing the stimuli (along with all other words in the story) a grand total of 40 times. Each training session required 40-60 minutes of laboratory time.

After concluding the training on the fifth day, subjects were given a short break and tested in a posttraining session[16] with the same words as used in the training sessions and, additionally, on a similar set of 12 pairs, the generalization set, not used previously in the experiment shown in Table 3 and Table 4 above. The generalization continua were tested in the same way as the training continua. Finally, the pretraining production procedure was repeated with two additions. First, a fourth contextual condition was incorporated. Orthographic segments from *The Selfish Recruit* were presented in randomized order interspersed with all other presentation stimuli. Each of the 74 segments was presented twice, increasing the production requirements by 148 phrases. Second, words from the generalization word sets were included in the isolation, carrier

[16] Since subjects were given posttraining testing on the same day as the final training session, it is not possible to strictly localize when subjects were trained, or the rate of training. That is, it is possible that training effects are the result of the last day of training only, rather than the accumulation of training over the five-day training regimen. Despite this detail, the results described below remain the same whether training was a long-term effect or the result of shorter exposure periods. Finally, production measures were collected continuously throughout the training, and, when analyzed, might bear on the rate and degree of training over time.

phrase, and coherent phrase context conditions (see Table 10 and Table 11 below for complete coherent phrases). Subjects produced these twenty-four words in each of three conditions four times each, for an additional 432 productions. Posttest productions totaled 1012 for each subject.

2.7. Methods: data design and analysis

Subjects performed a forced-choice two-alternative identification task in each of a pretraining, posttraining, and generalization sessions. The subject identified each presentation as belonging to the category of one of two words presented on the screen corresponding to either the voiced or voiceless member of a rhyming pair. Results were collected and coded automatically by condition.

A logistic regression analysis was performed on the data, and the odds ratios of each condition were calculated. From the resulting odds ratios, the unique equal-odds point on the VOT scale was estimated for each set of relevant factors in each condition. The estimated VOT value at the equal-odds point is the unique boundary location between voiced and voiceless for any given pair or condition. The set of VOT boundaries was then subjected to a repeated measures ANOVA and significance values computed.

Data were screened by subject and lexical item to ensure adequate task performance. Data was pooled for each lexical item across subjects, across training group, and by individual subject to confirm categorical selections were made under a variety of experimental conditions. Analysis showed consistent categorical identification including asymptotic leveling at stimulus extremes.

3. Results.

3.0. Results, introduction.

The research questions are: Do subjects trained on synthetically altered VOT shift their voicing boundary in perception after training? If such a shift occurs, does the boundary shift in the direction of training (either lengthened or shortened)? Is the effect different for high frequency words versus low frequency words? Finally, is there any effect on the VOT of words not used in the training portions of the experiment; that is, do subjects generalize in the direction of training.

In all cases, the dependent variable is the estimated VOT boundary, in milliseconds, calculated from the logistic regression analysis described above. Independent variables include training group (Group: lengthened, shortened), session (Session: pretraining, posttraining), lexical set (Set: training, generalization), and lexical frequency of both the voiced (Frequency Voiced: no, lo, hi) and voiceless (Frequency Voiceless: lo, hi) member of the pair constituting the opposing categories defining the boundary, for example, the pair *dime—time*.

3.1. Results, preliminaries.

The same materials were used for all subjects during all testing phases; the only variation was the stimulus (lengthened or shortened) during the training phase of the experiments, during which no perceptual data were collected. Subjects were assigned to either the lengthened or shortened Group without respect to testing procedures. The pretraining and posttraining Session used only the training Set while the generalization condition used a different, matched generalization Set. As discussed below, between-Set VOT values are not directly comparable, although within-Set values are appropriate for direct comparison because the lexical items are the same. Additionally, the generalization set was only tested *after* training, so there is no pretraining condition for the generalization set. Finally, Groups who are trained on both lengthened and shortened VOT do not violate standard orthogonality tests (W=.067, p=.181 and W=.148, p=.509, respectively) with respect to the variables Session, Set, or Frequency.

For all perceptual testing, subjects were presented with tokens from continua and asked to identify the presentation as either the voiced or voiceless member of the rhyming pair such as *take-dake*. Identification responses were collected for each continuum, and, as shown below in Figure 3, are plotted as proportion identified as 'voiced' (the ordinate) versus the VOT values of the presented token (the abscissa). In the figure below, a typical set of response functions is shown, in this case showing all of the identifications from one subject. The dashed line represents the pretraining identifications, the dotted line shows the posttraining identifications, and the solid line shows the generalization identifications. The equal-proportion boundary is shown by the horizontal dotted line in Figure 3 at 50% 'voiced' responses. Boundary locations are calculated as the point where a response curve crosses the equal-proportion point. Thus, the three values in the display below indicate the VOT at the intersections of each curve and the equal-proportion point. These values are interpreted as the unique boundary location for that pair. Similar response curves and boundary values were calculated for each continuum used in the experiment. The discussion below refers to boundary locations obtained in this way, in most cases averaged over the relevant variable.

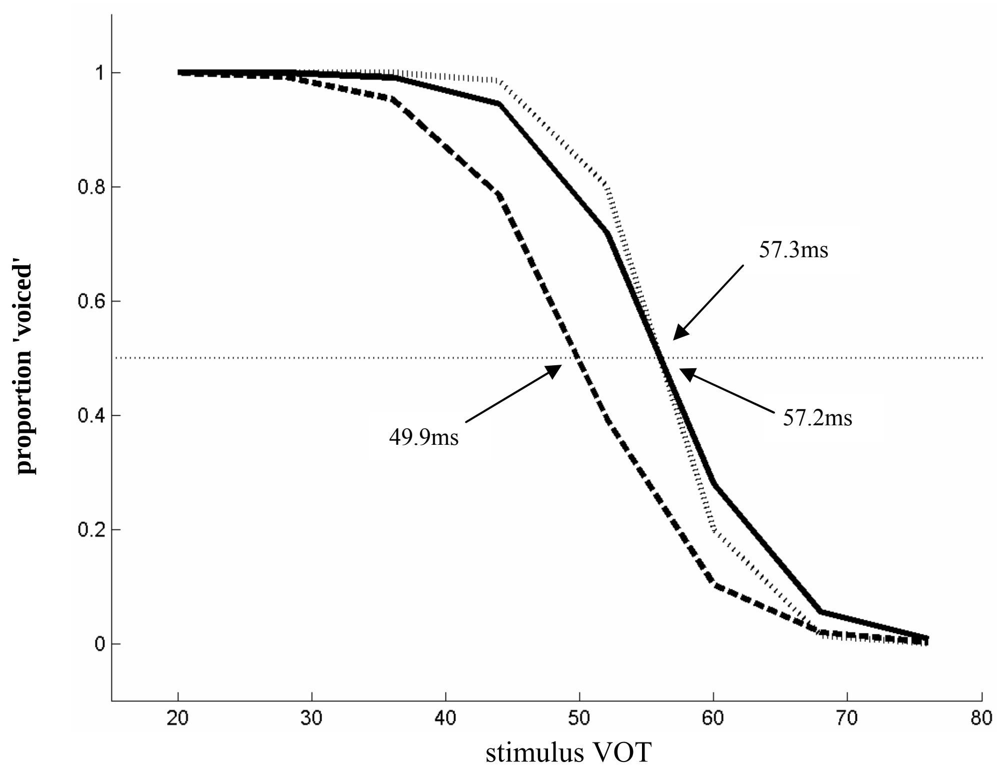

Figure 3. A typical identification function for all data from one subject in the lengthened condition. Identification responses are shown for pretraining (dashed line), posttraining (dotted line), and the generalization condition (solid line). The proportion of voiced versus voiceless responses is indicated on the ordinate, with the equal-odds point (0.5) shown as a horizontal line. The point that each response curve crosses the equal-odds point is the estimated boundary location, indicated as 57.3ms for the dotted line/posttraining, 57.2ms for the solid line/generalization, and 49.9ms for the dashed line/pretraining in the figure.

3.2.0. Results: Session, Set, and Group effects on boundaries.

There is a main effect of Session ($F(1,19)=30.64$, $p<.001$), indicating a difference between pretraining and posttraining boundary values. There is a main effect of Set ($F(1,19)=19.00$, $p<.001$) indicating a difference in the boundary values between the word sets used in the training portion of the experiments and the generalization condition. The main effect of Group did not achieve statistical significance ($F(1,19)=3.39$, $p=.0813$). This lack of significant difference between the lengthened and shortened training groups, however, is not likely to be an important result since there are a number of variables and associated (interaction) effects. Further, a main effect of Group is not necessarily predicted, since Group is arbitrarily assigned for both the pretraining and generalization conditions. That is, there is no experimental or expected difference in the pretraining condition since all subjects performed identical tasks. To more closely examine relationships among variables where differences are expected or predicted, the relevant interaction effects are given below.

The lengthened Group posttraining boundary is longer than the pretraining boundary for that Group as well as the boundaries for both Sessions of the shortened Group. The Group trained on lengthened VOT has a longer posttraining than pretraining boundary ($F(1,9)=76.37$, $p<.001$), but the same within-Group effect is not found for the shortened Group ($F(1,9)<1$, $p=.409$). There is also a between-Group posttraining effect such that the boundary is longer for the lengthened Group than for the shortened Group ($F(1,19)=9.61$, $p<.006$). The effect of lengthened training is a boundary increase of 3.3ms within-Group and 2.1ms between-Group. Figure 4 below shows the mean VOT boundary values for both Groups and Sessions. Values are given in the display at means, and error bars throughout are reported at 1 standard deviation.

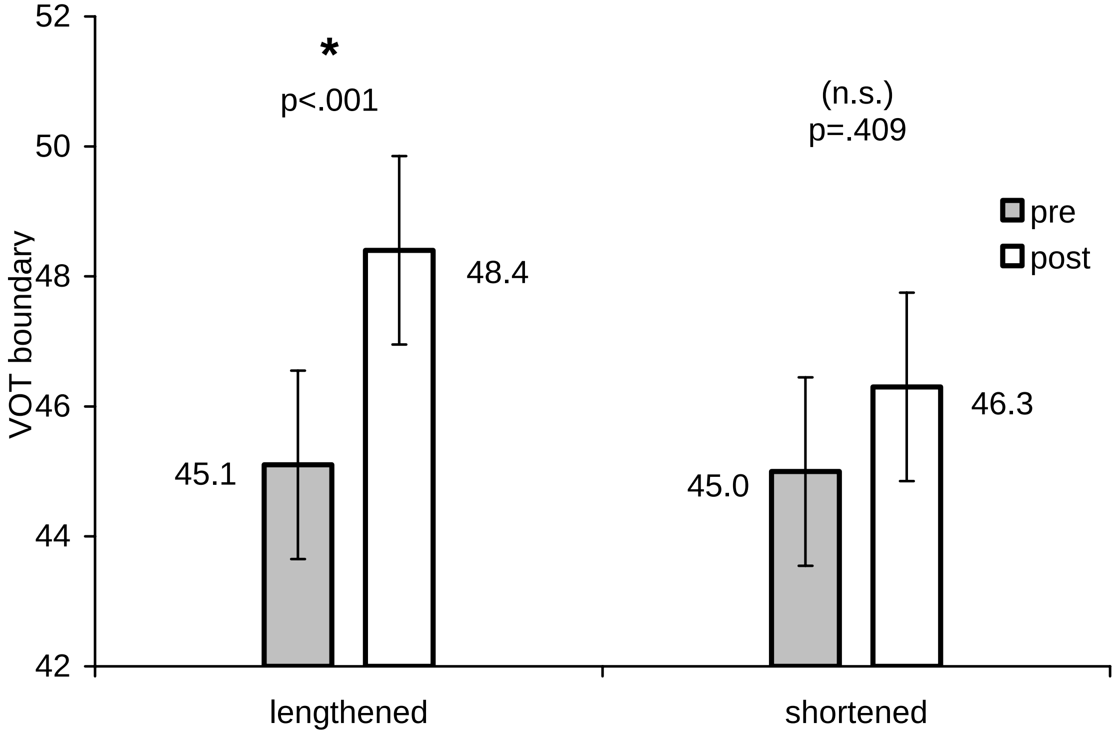

Figure 4. Main effects of training. Pretraining and posttraining VOT boundary vales for subjects trained on lengthened tokens and subjects trained on shortened tokens. Subjects trained on lengthened tokens show a 3.3ms VOT boundary shift in the direction of training after training. Subjects trained on shortened tokens do not reveal a statistically significant effect of training.

There is an asymmetry in the design mentioned above resulting in no pretraining boundary estimate for the generalization set of lexical items. Between-Set comparison of absolute boundary values is inappropriate due to the possibilities of unexamined factors entailed in each Set's lexical items. However, the Set variable is nonetheless useful to identify within-Set category differences at constellations of other variables (that is, three-way interactions). Although there is a main effect of Set, this effect is not important to the present discussion because that effect is a reflection of between-Set differences. The question with respect to the generalization condition is whether subjects successfully trained (in any direction) extend that training effect beyond the specific words trained on.

Since direct comparisons of absolute values are impossible due to the different words used in each Set, this carryover effect can be demonstrated by showing that a successfully trained Group performs differently on the trained-on lexical items than on new lexical items.

Specifically, since there was a clear between-Groups training effect in the training Set, subjects can be considered to extend this training if they show a (similar in direction and magnitude) difference in a new Set. On the other hand, if differently trained Groups do not perform differently on a new Set, then the effects of training cannot be considered extended but rather the successful training effects must be considered to be restricted to the trained lexical items.

There is a three-way interaction between Group, Session, and Set ($F(1,19)=4.49$, $p=.0473$), with pairwise comparisons revealing the effect of training Group is not significant, neither within the pretraining Session ($F(1,19)<1$, $p=.976$) nor within the generalization Session ($F(1,19)<1$, $p=.924$), but is significant in the posttraining Session ($F(1,19)=11.01$, $p<.004$). All values are given below in Figure 5.

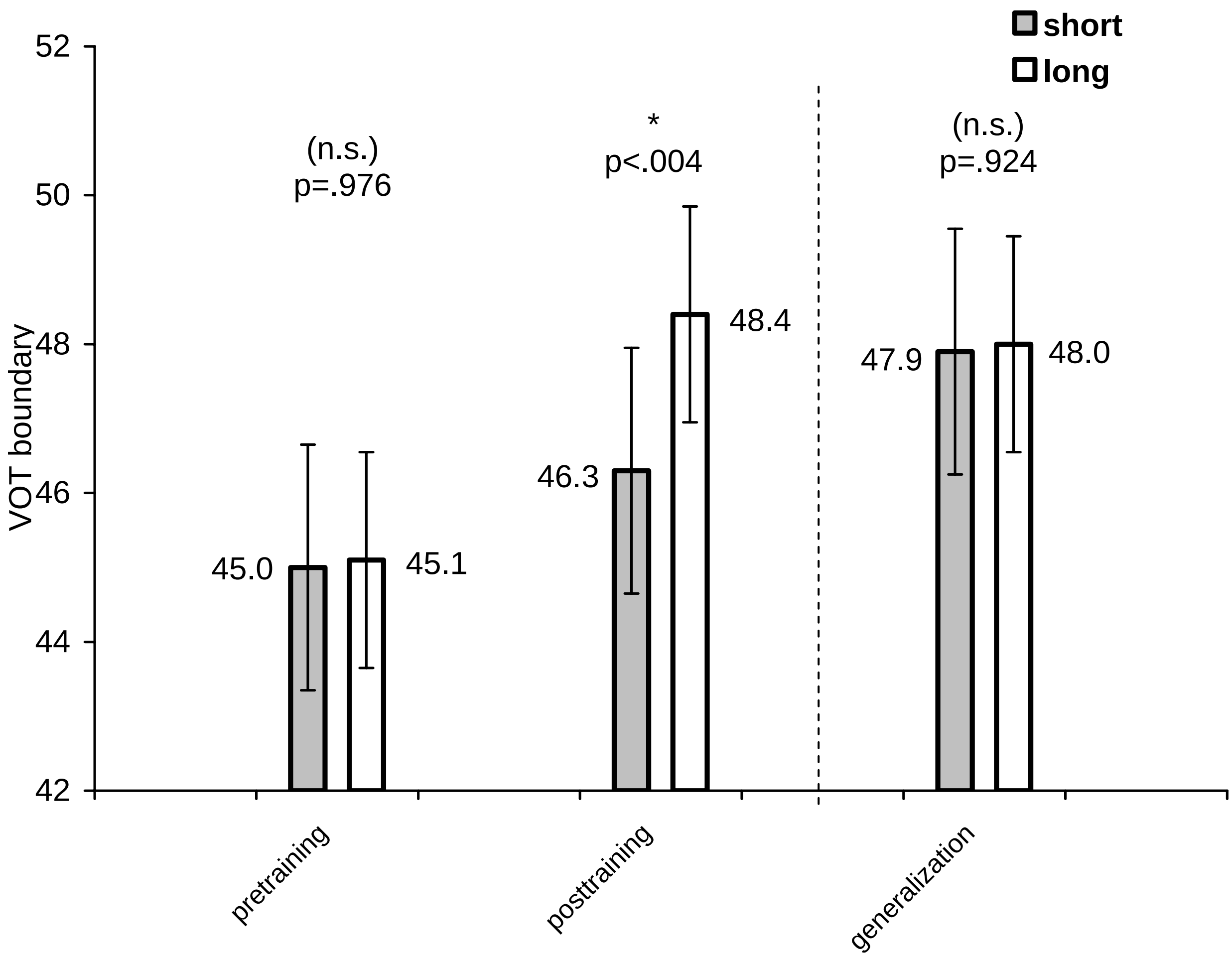

Figure 5. The interaction effects of Group, Set, and Session. Each pair of bars indicates the shortened training group on the left and the lengthened training group on the right. The voicing boundary values given reveal no significant difference between the two training groups in the pretraining condition (when Group affiliation is entirely arbitrary) or in the generalization condition in which the target words were previously not used in the experiments. The only significant differences among these variables is between the lengthened and shortened group after being trained. Although significance is given only for the pairwise comparison between the two groups at the posttraining condition, the lengthened group is also significantly different from both VOT boundary values in the pretraining condition. The dashed line delimiting the generalization condition indicates the different word set used for that condition, and, as a consequence, the inability to directly compare the actual boundary values of that condition to values obtained using a different set of words.

3.2.1. Results: lexical frequency

The present boundary estimates are calculated based on subjects' word identification of a specific presentation pair differing only in the voicing value of the initial stop consonant. In each case of boundary identification, the subject is asked to choose if an auditory presentation belongs to the voiced category, for example *deal*, or the corresponding voiceless category, *teal*. Lexical frequency is independently controlled for each voiced and each voiceless member of a pair, Frequency Voiced (no, lo, hi) and Frequency Voiceless (lo, hi). Results pooled across Group and Session are shown in Figure 6 below. Notice that in the following Figure 6 all the relevant data is actually repeated on each side of the dotted line delimiting the display. At each frequency level of each variable, the full range of levels is symmetrically represented from the opposing category. For example, for the level 'NO' in the voiced frequency variable, both 'LO' and 'HI' are equally represented.

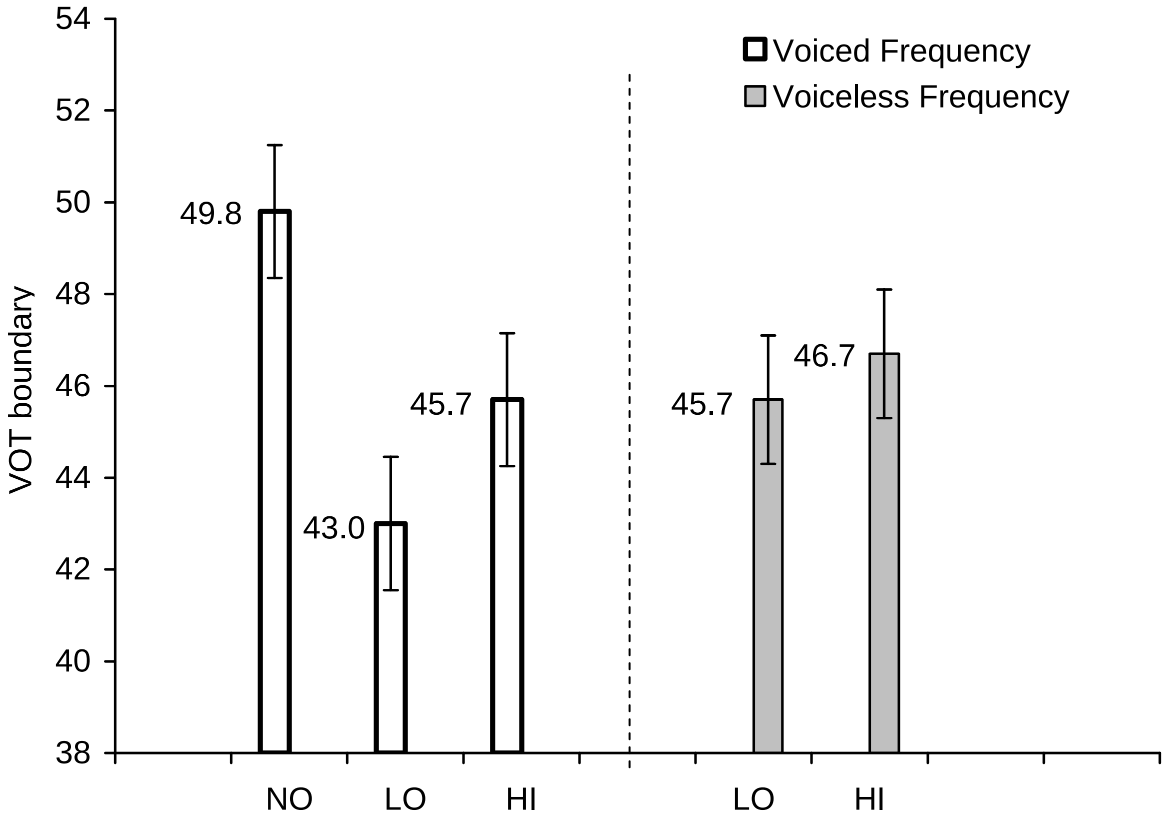

Figure 6. VOT boundary values by lexical frequency specification. The lexical specification of the voiced word is shown to the left of the dashed line, and the specification of the voiceless on the right. This display shows the VOT boundary location for all training continua used in the experiments. Since all continua must have both a voiced and voiceless pair-member, all training continua are represented twice, once as the frequency of the voiced pair-member (left of the dashed line) and again as the frequency of the voiceless pair-member (right of the dashed line).

There is a main effect of Frequency Voiced ($F(2,19)=66.32$, $p<.001$) but no main effect of Frequency Voiceless ($F(1,19)=1.05$, $p=.317$). The pairwise comparisons among the three-leveled voiced frequency variable further reveal a three-way differentiation in which each level is significantly different from the other two. Next, Frequency does not interact with training Group. As shown in Figures 7 and 8, both training groups perform similarly with respect to Frequency.

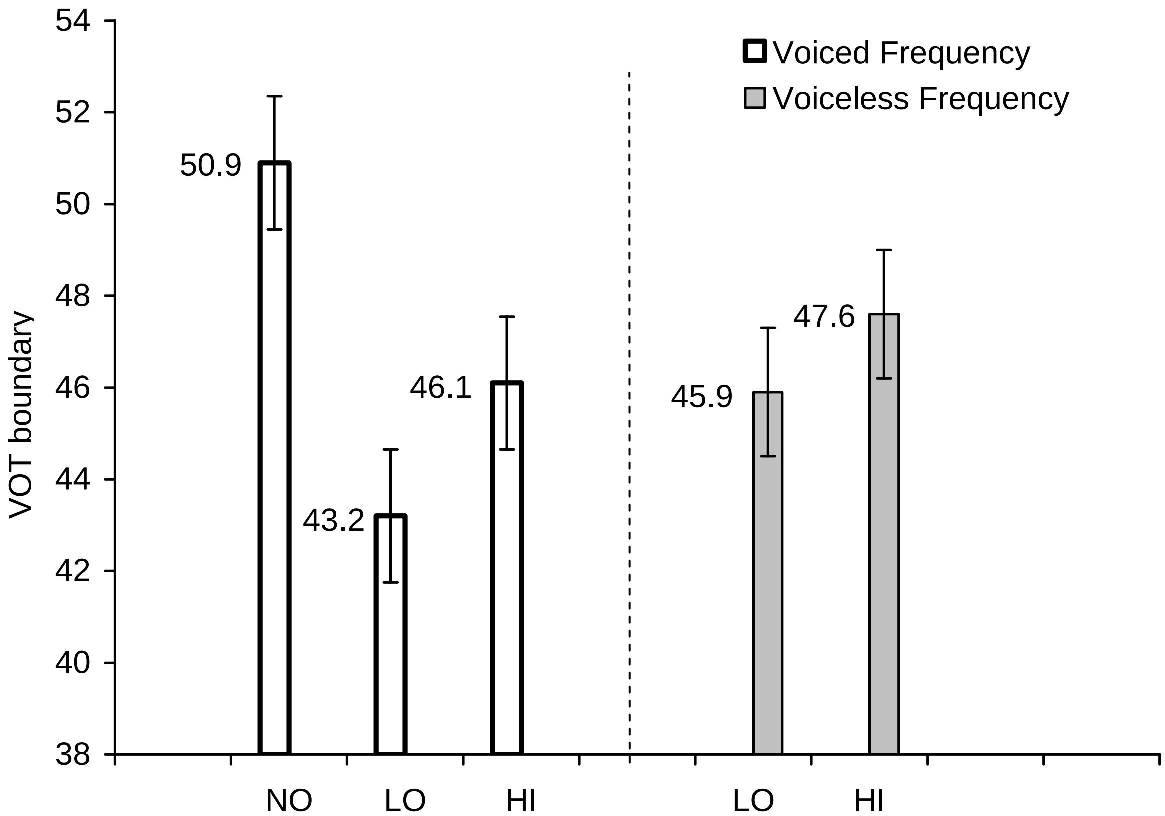

Figure 7. Voiced and voiceless VOT boundary locations for lengthened training group, training lexical set only.

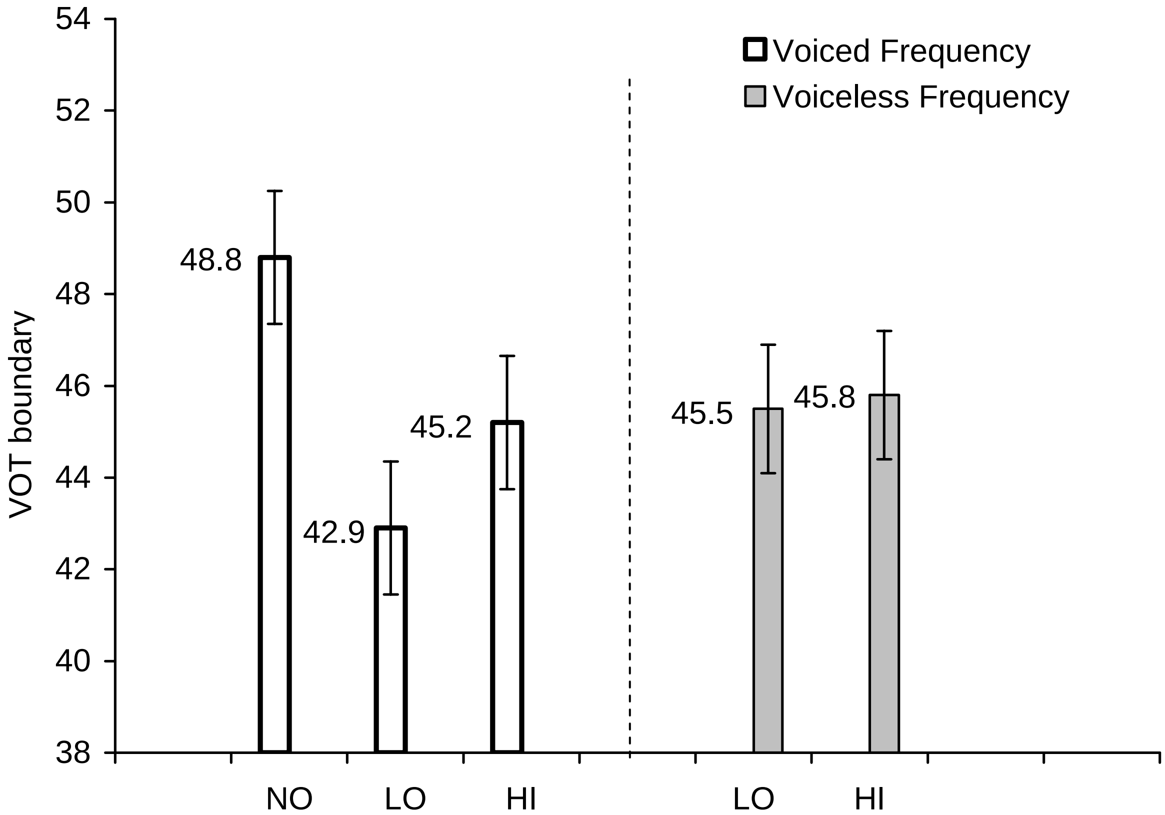

Figure 8. Voiced and voiceless VOT boundary locations for shortened training group, training lexical set only.

Statistical tests reveal significant differences within the Frequency Voiced, but not within the Frequency Voiceless categories for both training groups. That is, the pooled data is conceptually and statistically similar to the data organized by Training group.

However, as shown above, there is an interaction effect where the lengthened training Group showed an effect between the pretraining and posttraining Sessions. To investigate the possibility of a three-way interaction between Frequency, Training Group, and Session, Figure 9 and Figure 10 below show each training Group in different panels and each Frequencies in paired bars, the left indicating pretraining and the right posttraining.

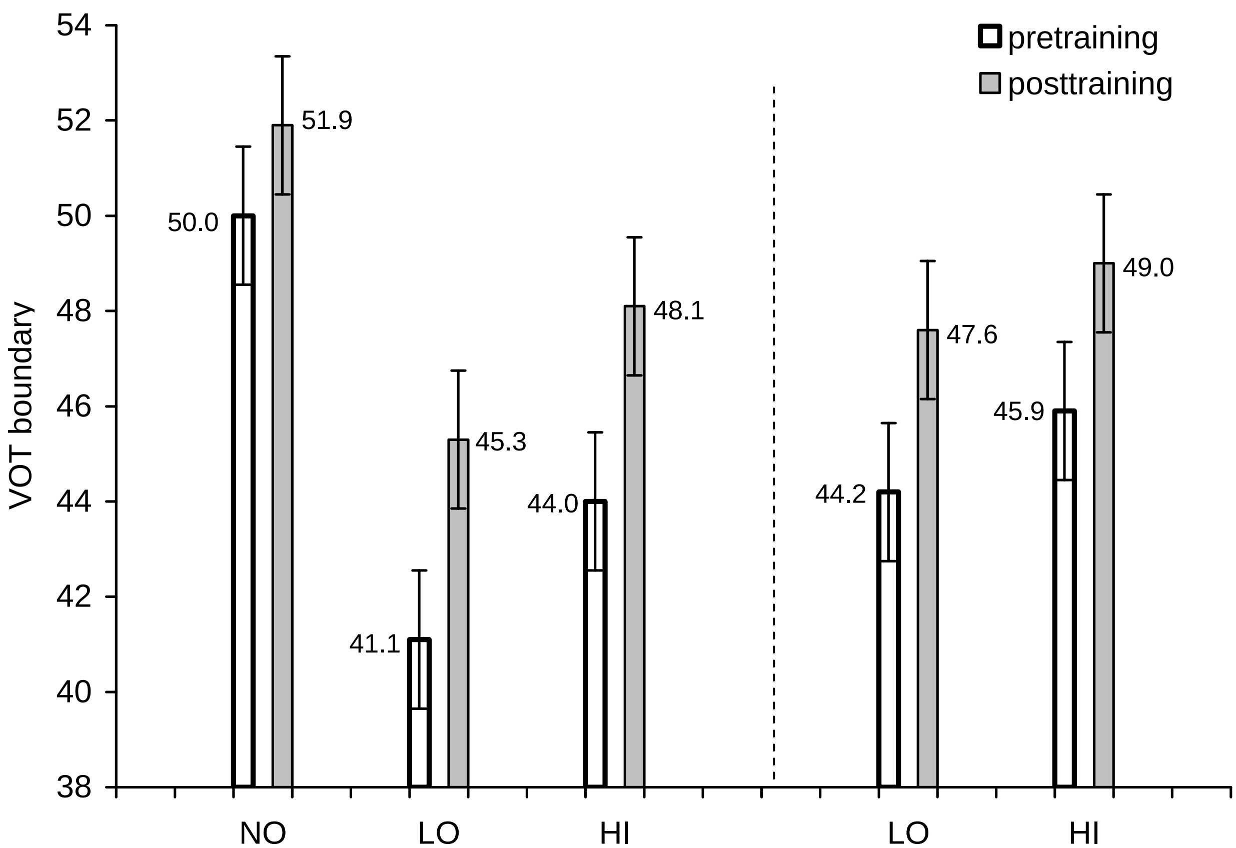

Figure 9. Boundary VOT values for the group trained on lengthened tokens is shown. The display is arranged similar to displays above, indicating the voiced series (no-, low-, and high-frequency) to the left of the dotted line, and the voiceless (low- and high-frequency) to the right. Each pair of bars shows the pretraining boundary value on the left and the posttraining boundary value on the right. Only forms from the training set are shown here. This display clearly shows the robust training effect across frequency specification.

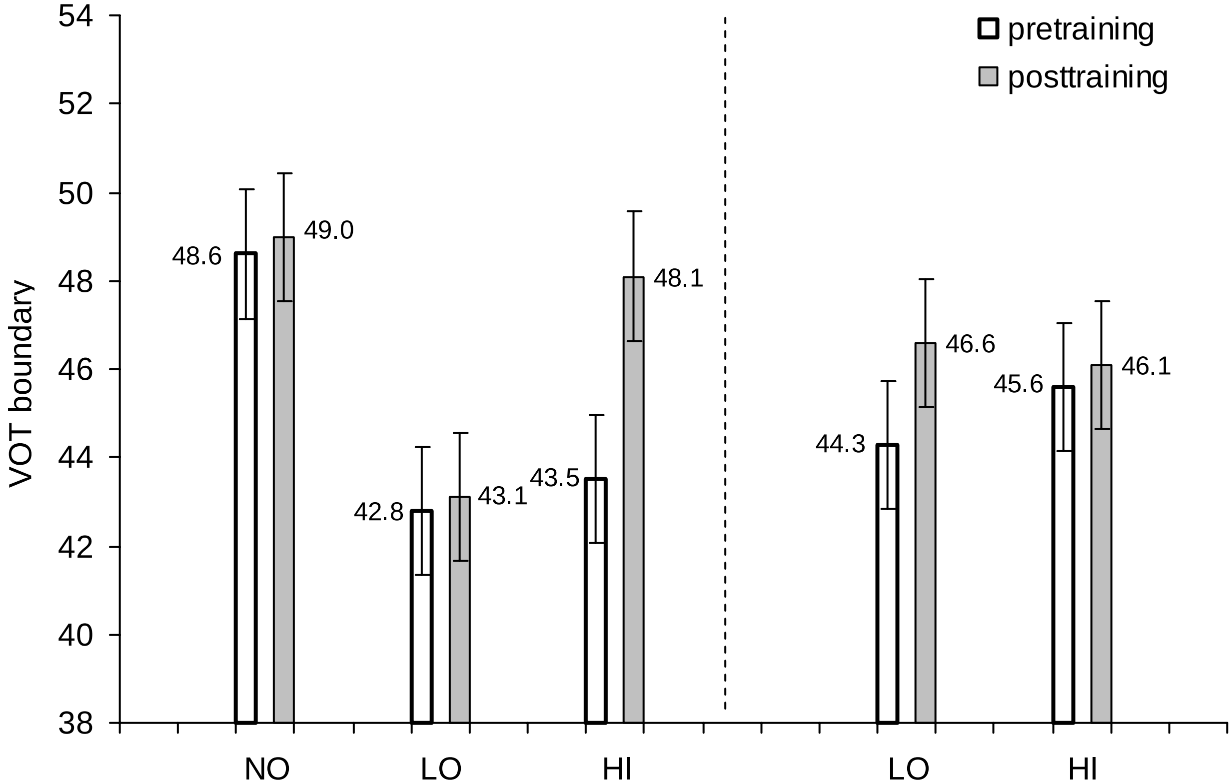

Figure 10. Boundary VOT values for the group trained on shortened tokens is shown. The display is arranged similar to displays above, indicating the voiced series (no-, low-, and high-frequency) to the left of the dotted line, and the voiceless (low- and high-frequency) to the right. Each pair of bars shows the pretraining boundary value on the left and the posttraining boundary value on the right. Only forms from the training set are shown here. This display clearly shows the lack of consistent training effect across frequency specification.

There is not a three way interaction between Group, Session, and Frequency. The main training effect for the lengthened training group, however, is clearly visible in Figure 9. In this display, importantly contrasting with Figure 10, each posttraining boundary value is significantly longer than the paired pretraining value. Again, the direction and magnitude of training is evident at all Frequency types. The same relationships and consistencies are not found the shortened training Group.

4. Discussion.

4.0. Overview of discussion.

This chapter first briefly reviews the basic results from the present work. One primary result, the posttraining category shift, is given attention after a review of previous work on similar boundary shift research in the literature. Another main result, the effect of lexical frequency, is then considered in detail before a short discussion of the role of interactions among factors. The results are then considered more broadly, in terms of the two theoretical models described earlier, the traditional model and the exemplar model. Finally, a number of implications from the present work conclude the discussion.

4.0.1. Summary of main results.

The present work shows that listeners' linguistic representation of words in memory might include factors such as lexical- and frequency information. These features are shown to be highly specific, not generalized throughout the lexicon, and subject to specific linguistic experiences that affect their representation in memory. These results present a theoretical difficulty to the traditional, abstractionist language model but are consistent with episodic language models.

4.1. Plasticity of linguistic representations.

This section pursues evidence in the literature of the plasticity of linguistic representations over time under various contextual conditions. Before examining specific evidence of category plasticity (e.g., a category boundary shift), it is appropriate to briefly discuss the notion of change over time. Described in detail below, the changes relevant to the present discussion of plasticity are framed in terms shorter than traditional historical or diachronic language change while simultaneously large enough to be considered squarely a linguistic issue with implications for the language system, rather than only a mechanical, or performance issue. That is, categorical plasticity is considered at a time scale from a few hours to a few days. Specific categorical plasticity reported in the literature is then discussed.

4.1.0. Framing the time scale of linguistic change in speech perception.

The traditional view of the lexicon including such parts as category centers and boundaries supposes that the lexicon is practically stable at any given time in a language user's history, but may change given a long enough time frame. Much previous work has addressed historical or diachronic change including the dynamic nature of the lexicon over relatively long periods of time such as years or generations (Brugmann, 1886; Bloomfield, 1933; Hock, 1986). The issue normally thought to be addressed in such works is the state of an entire language system as it is used in an active language community. Of course any change is manifested in individuals, but the scale of change is assumed to be at the level of shared language in its social context. Changes of this sort are normally thought to change the systematic grammar of the language and are thus irreversible with talkers incapable of returning or reverting to older forms. Indeed, if the time frame is as long as a generation, changes are actually acquired as native and there is no possibility of a return to another, older form since there would be no representation of that form at all.

The present work, by contrast, is interested in representational change within a much shorter time frame, in terms of an individual's memory for words and language at intervals normally not examined at an interval more than a week or so (Goldinger, 1996) and in some cases on the scale of only a few hours (Shockley et al, 2004). In this sense, the present work is more about language (category) learning or training than about language change in the usual diachronic sense used in linguistics.

The time frame of plasticity considered here, between hours and days, is important for a number of reasons, including the issue of access to linguistic representations. Since the present study depends on lexical access, for example, to the lexical frequency of a form, the time frame must be long enough for lexical access to occur. As Norris et al explain, "if lexical information can play a role in modifying phonetic categorization over time this would involve higher-level (lexical) knowledge" and the problem is then claimed to lie within the linguistic domain rather than falling into a lower-level domain such as the motor domain, for example (2003: 205). On the other hand, the experimental time frame cannot be so long as to introduce a wide variety of unknown influences or permit further changes or reversions (for example, due to representational decay consistent with exemplar models). For example, given these concerns, Goldinger (1996) reports that recall performance declines over time such that immediate recall performance is better than recall performance after a week while other reports suggest that a delay of about 12 hours has little to no effect on affected representations (Eisner & McQueen, 2006). Presumably linguistic representations are stronger after shorter delays than after longer delays, at least up to a week. Notice that although representational strength declines over the course of a week (a time-related decay function), Goldinger still reports a detectable effect of representational plasticity persistent after a week when subjects are asked to judge which forms are more similar to each other.

The delay between experience and a detectible effect, however, obviously has a lower limit, below which effects are not detectable. In terms of speech processing, Pisoni claims that "perceptual processing is completed within 250ms" and that such processing is "absolute and binding" by such time (1973: 259). Balota and Chumbley (1984, 1985) examined the interaction of lexical frequency and enforced response delays in a reading aloud task, finding response latency decreasing across delay intervals up to about 500ms (after which, the latency appears to level off). This suggests lexical access is achieved by at least 500ms. More recently, exploring vowel production and the role of lexical access utilizing Balota and Chumbley's methods, Munson (2004) contrasts an 'immediate-response' condition in which stimulus presentation and the signal to respond are simultaneous with a 'long-delay' condition in which there is a 1000ms delay between stimulus presentation and the response signal. He describes the 'immediate-response' condition of lexical access as 'stressed' and the 'long-delay' condition of lexical access as 'facilitated'. Under these conditions, he found greater vowel dispersion and reduced response latency in the long-delay/facilitated condition as compared to the immediate-response/stressed condition. The response latency for the long-delay condition was about 91ms shorter than for the immediate-response, which is interpreted "to reflect the additional time required to conduct lexical access in the immediate-response condition" (in press). For purposes of the present work, it is important then to consider changes long enough to comfortably assume lexical access (i.e., greater than 250-500ms) but not too long as to obscure potential effects or become impractical, that is, less than one week, but even shorter might be better.

Additionally, category shifts in the present work are most likely temporary changes at the level of plasticity in the individual lexical items. They are not changes in the phonology (as described below) resulting from direct experience. Further, those representations are assumed by exemplar theory to decay or revert back to the previous representations at a certain rate. The plasticity (lexical category shifts) relevant to the present discussion include those changes that are the result of targeting specific representations rather than the phonological system, as discussed below as the lack of generalization.

So, with preliminary comments concerning the possible time-course implications of plasticity and lexical change considered immediately above, the following discussion presents research that has been conducted under the assumption that categories and associated boundaries are indeed plastic. The literature considers a number of conditions under which linguistic representations can be explicitly altered through a variety of exposure techniques. Since plasticity is related to the time course of exposure or experience, the review below examines the literature addressing category shifts at progressively longer time frames, from on-the-fly effects to perceptual learning after delays exceeding 12 hours and an intervening sleep cycle.

4.1.1. Verbal transformation modifies category perception at a very short time delay.

Verbal transformation (also referred to as resyllabification), interpreted here as categorical perception shifts, occurs after repeated exposures to the same stimulus. For example, when a stimulus syllable such as *life* is repeated in rapid succession, observers report the auditory illusion of perceiving transformations between *fly* and *life* (Warren & Gregory, 1958; Warren, 1970). This result has been replicated under a variety of conditions and contexts (Stetson, 1951; Warren, 1961; Sato et al, 2006; de Jong, 2001a, 2001b, 2003; de Jong et al, 2004). For the present discussion, several facts concerning this research are important. First, the transformation is reported only under fast repetition conditions. Second, the transformation is only a product of the task and does not persist beyond the immediate task into later perceptual tasks. That is, if a listener reports perception of a particular percept, say *fly*, when the test is completed, the listener immediately returns to normal perceptions of *fly* and *life* upon re-testing. Verbal transformation does not apparently influence the representation of categories (although it is argued to access representations (Sato et al, 2006)). Third, verbal transformations only occur with syllable-sized units, not larger constituents that include linguistic context. Taken together, these facts are important to show that the speech perception mechanism is very fast and capable of on-line plasticity.[17]

4.1.2. Selective adaptation modifies category perception.

Perception of linguistic categories also shifts as a result of selective adaptation to a stimulus. Selective adaptation is achieved by repeating an adapting stimulus many times in a short period before testing a specific, categorically related test stimulus. Eimas and Corbit (1973) first obtained linguistic identification functions for apical and bilabial

---

[17] It should be noted that in this case there is no claim of representational shifting or categorical alteration of any sort. Perceptual alternation is achieved across some sort of organizational boundary (such as across the voicing boundary (de Jong et al, 2004) or across a syllable boundary as indicated by lexical identification (Warren,1961)), and it is normally considered a form of on-line access rather than a persistent category shift.

series of stop consonants in an unadapted, baseline state, then presented subjects with the repeated adapting stimulus from a portion of the voicing continuum preselected as a good within-category exemplar. An identification task similar to the baseline task was repeated after the adaptation exposure phase, and pre- versus post-identification functions were compared for difference. It was found that the boundary location was shifted in the direction of the adapting stimulus after the adaptation exposure, indicating the plasticity of the category. Notice the fact that the boundary shifted in the direction of the adapting stimulus means that the response function for the category containing the adapting stimulus was reduced. That is, after adaptation ambiguous tokens (from the boundary region) were more likely to be identified as belonging to the category *opposing* the adapting stimulus. Eimas and Corbit (1973) report a mean adapted boundary shift of 8.0ms, with the effect stronger for the group of subjects exposed to a voiceless adaptor than for the group exposed to a voiced adaptor (10.0ms and 6.1ms, respectively). Similar results were obtained by Tartter and Eimas (1975), Cooper (1974, 1979), and Ohde and Sharf (1979), although in those studies the degree of shift was slightly less (about 3ms, 5ms, and 6ms, respectively) as compared with the original seminal study by Eimas and Corbit. Selective adaptation results were also obtained in other phonetic domains such as amplitude intensity (Ohde & Sharf, 1979), place of articulation (Cooper, 1979; Pisoni & Tash, 1975), and adaptation of speech to non-speech stimuli (Samuel & Newport, 1979; Kat & Samuel, 1984).[18]

[18] The explanation offered by selective adaptation for *why* this occurs is that certain (feature or acoustic) detectors become fatigued and are rendered less able to attract ambiguous tokens into that category, which then results in identification into the opposing default category. Of course, this line of reasoning assumes or depends on a binary category distinction for which a default can be assumed, a gradient feature that can be varied, interpretation that a single feature or feature detector is responsible for the adaptation, and the lack of other reasonable explanations consistent with the data and phenomena (Rosen, 1979). Serious theoretical problems with selective adaptation have been explored by many including Remez, Cutting, &

A conceptualization of the experiments and typical results are idealized in Figure 11 below. In the top panel, response frequency is represented on the ordinate, and perceptual category identification (*F*)[19] and a corresponding gradient auditory dimension is represented on the abscissa. For example, supposing representation of the voicing contrast, the abscissa represents continuous values of VOT correlating strongly with degree of voicing, upon which the feature (*F*) is mapped corresponding to the categories *voiceless* (+*F*) and *voiced* (-*F*). The response functions are maximal at each category center, and decrease moving away from the category center either toward the continuum extremes or toward the estimated boundary value (defined as where the response curves cross). Similar normal response curves of equal size for each category is assumed in this model (Diehl, 1981; Miller & Volaitis, 1989). The middle panel shows an adapting or training stimuli in the approximate center of the +*F* category (although category selection is arbitrary). No response was recorded during the adaptation/training phase, but the training stimuli are defined to be centrally located within the chosen category. Finally, in the bottom panel, hypothesized post-exposure response functions and boundary locations are shown. With the adapting stimulus presented in the +*F* category, the response function to that category after exposure is weakened (suggested by a depressed, smaller response curve), and the subsequent right-shifted boundary is observed. The post-exposure result here is that tokens in the boundary region, i.e., ambiguous tokens, are

---

Studdert-Kennedy (1980) and Diehl (1981), and the practice has, for the most part, been abandoned in speech research since the 1980s. The purpose of the present discussion is not to explore the value of selective adaptation or even its theoretical implications on issues such as the nature of lexical access during speech perception. Rather, the present discussion is intended to point out the time-course of multifarious, on-line flexibility of speech categories and the direction of category shifts after experimental exposures.

[19] In this case, the category *F* is used as shorthand for what a perceiver identifies as belonging to some (binary) value of *F*, which may be understood here to be the recognition of any arbitrary, minimally different category or feature that minimally differentiates such categories.

more likely to be categorized as –*F* after exposure than before since the boundary has shifted to make the –*F* category larger.

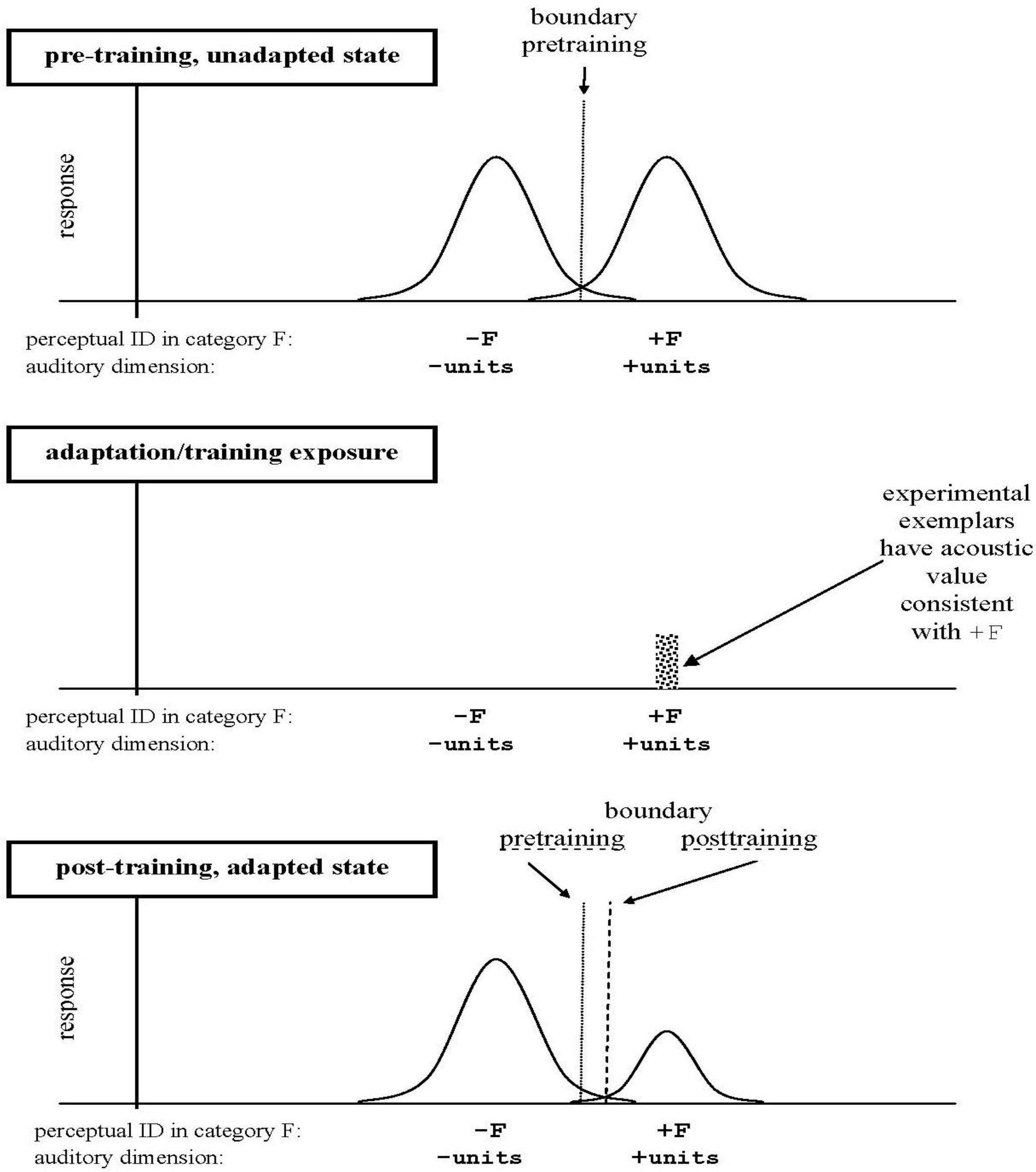

Figure 11. Schematic response curve shift after selective adaptation. The top panel shows the hypothesized baseline state for two categories defined along an auditory dimension (in units) and a binary perceptual category labeled 'F'. The middle panel shows the training phase in which a large number of +F tokens are presented to a subject. The bottom panel shows the resulting response curves after training, in which the trained-on category response function is reduced. This reduced response is interpreted as a (rightward) boundary shift toward the reduced category, indicating the trained-on category is smaller in post-training testing than in pre-training testing.

The general result of selective adaptation is that the training feature is interpreted as weaker after training. The present work, however, found just the opposite result in which the response function to the training feature, lengthened VOT, is stronger after training. The present study successfully trained subjects on lengthened VOTs of target words, and posttraining testing revealed that the voicing boundary shifted, making the trained category 'stronger' (or larger in terms of the auditory scale).

Despite the selective adaptation and present work offering description of the plasticity of representations, the results from the present study call into question the robustness of the selective adaptation studies. There are, however, two important differences between selective adaptation and the present work. First, as practiced, selective adaptation requires a very large number of exposures, often hundreds of adaptation trials (Pisoni & Tash, 1975), using a highly unnatural non-sense syllable out of linguistic context as the adapting stimulus (Eimas & Corbit, 1973; Cooper, 1974, 1979), then a test phase immediately following the adaptation procedure. These experimental features present obvious problems of ecological validity generally, but also differ from the present study in which training was performed with a small number of stimuli, actual words, highly natural stimuli, under a variety of linguistic contexts.

The second important difference is the apparent necessity of a short time window between the adapting phase and the testing phase for selective adaptation, but up to week-long testing for the present work.[20] The literature on selective adaptation does not

[20] As mentioned earlier, it is possible that the effects reported here are the result of only the last day of training. This possibility, however, was not directly tested, and since testing measures from only the first and last days were recorded, any changes reflect differences that could arise any time in the five days of the

thoroughly consider the possible persistence of an adaptation effect beyond the immediate post-adaptation test phase, although selective adaptation shifts have been reported when tested two seconds after adaptation, slightly less shifted after 5 minutes, and no longer detectible after a 30-minute interval between adaptation and testing phases (Harris, 1980). In most cases the selective adaptation procedure simply tacitly demands compression of experimental phases as a methodological practice adopted under the assumption that it would increase the likelihood of obtaining effects (Rosen, 1979). Delays up to 24 hours were imposed in order to replicate experimental effects with different adaptors, assuming the effects of the previous day had no influence (Eimas & Corbit, 1973). The present study, however, demonstrates that plasticity is observed when strict experimental compression is not maintained, and even raises the possibility that there might be training effects that persist for longer than was assumed in the selective adaptation literature. The present results are in concord with a variety of more recent results on the persistence of the plasticity of linguistic categories over time (Eisner & McQueen, 2006).

4.1.3. Categories shift in response to shadowing or immediate recall tasks.

Another line of research demonstrating plasticity in speech categories is research on shadowing. In shadowing tasks (or mimicry or imitative-response), subjects are given an auditory presentation and asked to repeat that presentation as quickly as possible. Although procedures and tasks vary according to different studies (for example, eliciting 'choice response' or 'simple response'), shadowing research consistently shows that

---

experiments. It might have been possible to retest the target training words on the last day before training, but then the amount of exposures to target words would also necessarily change.

language users are capable of fast and reasonably accurate shadowing. In a series of separate reports, subjects were asked to shadow stimulus vowels or consonants that would unpredictably and spontaneously change during presentation. Subjects recognized the change either by producing a pre-specified articulation ('simple response') or producing the quality of the stimulus change ('choice response') (Porter & Lubker, 1980; Porter & Castellanos, 1980; Fowler et al, 2003). For simple responses, subjects indicated the existence of some change while for the choice response condition, subjects were required to specify precisely which change occurred (i.e., the quality of the change). Subjects performed this task with relatively high accuracy and speeds comparable with other, simpler acoustic reaction time tests.

These studies tend to support a model in which perceptions can directly, quickly influence productions, and, under certain circumstances, these influences enter into the speech mechanism "as continuous adjustments, tuned to characteristics of the ongoing signal" (Porter & Lubker, 1980: 1349). This is important to the present work because, as described below, access to certain lexical properties such as frequency must be established in the listen-and-repeat task cycle performed by subjects in the present work. Establishing empirical support for this lexical access, Porter and Lubker further argue from comparing simple and choice response data that subjects are able to perform choice response and simple response tasks at about the same lag (less than 250ms for all studies, but Fowler et al, 2003 was much faster), despite choice response necessitating a level of processing assumed to include access to deeper (linguistic or lexical) memory. Porter and Lubker argue that since the reaction time data is assumed to require lexical access in the delay condition, and, critically, that the latencies are comparable in both the delay and immediate response conditions, lexical access can be assumed in both conditions. That is, they argue (albeit from negative results) that since one condition requires the assumption of lexical access, and the results of both conditions are indistinguishable from one another, they must both be compatible with a lexical access account. With respect to the present discussion, this is evidence that language users closely monitor stimulus cues during shadowing tasks, and, further, that subjects show detectable evidence of fast, accurate, online access to lexical properties of words. In an exemplar model, there may be many features stored in memory, and, as a result, those features might be accessible on-line to dynamically affect the representation of those categories in memory. Thus, the potentially large number of on-line influences able to affect the representation of speech could help to explain the results of current study.

Following up on exemplar modeling consistent with a large number of features in memory, Goldinger elicited both immediate- and delayed-shadowing productions (1996, 1998). A panel of judges later listened to both shadowings evaluating which sounded more like the stimulus in an AXB discrimination task. The immediate shadowings were judged more like the stimuli than the delayed shadowings (based on faster reaction times). Furthermore, lexical frequency and number of repetitions of target words was considered, and results indicate that reaction latencies decrease as frequency increases, and reaction latencies decrease as number of repetitions increases. These differences are interpreted to indicate that subjects access some part of the lexical representation of the target words, and, since tokens with more experiences were responded to faster (either in terms of lexical frequency or terms of experimentally controlled repetitions), Goldinger argues in favor of a rich memory, exemplar model of the lexicon in which experiences of particular lexical items are encoded into the representation of that lexical item. In the delayed condition, subjects waited a few seconds before shadowing the stimulus presentation, and performance declined (latencies increased). It was argued that the differences between immediate and delayed judgments are due to subjects performing more lexical access during the delay interval. It is argued that this access allows subjects to compare the stimulus with lexically stored exemplars, and, after such lexical access, the resulting production is less similar to the stimulus because it is influenced by stored lexical representations (which presumably differ in meaningful ways from the stimulus). Overall, Goldinger's experiments show that as a result of specific experiences—in this case, shadowing—subjects change their productions to more closely match the stimulus in ways detectable to listeners. He claims that those changes are lexical (and not, for example, motor or automatic) and thus have influence on the representation of the relevant forms, indicating a plasticity of category representation.

Reporting similar results, Shockley et al (2004) replicated the major result of Goldinger (1998) showing that subjects judged shadowed repetitions more similar to a stimulus than baseline (non-shadowed) productions. In addition to replicating Goldinger's results, Shockley et al manipulated stimulus tokens so that the VOT was about twice its natural duration (a mean absolute increase of about 68ms), which were "noticeably breathier than the original productions... [but] did not sound unnatural" (2004: 425). Subjects were then asked to perform a shadowing task on these synthetically lengthened stimuli. The results bear on both production and perception. In terms of production, the words shadowed after a lengthened stimulus were longer than the same words in the baseline condition (by about 12ms), indicating that subjects affected their productions to be more similar to the stimulus, again showing flexibility in category structure. In terms of perception, a panel of independent judges evaluated the shadowed productions as more similar to the artificially lengthened stimuli than to the baseline productions.[21] Interestingly, the possibility that these results were artifactual of overall word lengthening (possibly due to decreased naturalness or goodness) was dispatched after failing to find a significant effect of word length in addition to the lengthening directly resultant from the VOT lengthening. That is, for both the productions of shadowed words and the perception-only tasks performed by additional subjects, subjects were able to home in on specific, sub-phonemic detail of the signal and use that detail later in recognition tasks. When subjects were asked to produce shadowed words, a group of judges were able to successfully identify the shadowed productions as more similar to the stimuli than other, non-shadowed productions. The Shockley et al results indicate that language users actively adjust the location of categories based on exposure to specific exemplars, although for both the Goldinger and Shockley et al studies, the effect of experience was greatest when it was temporally closest to the evaluation or production. In terms of the present results, these shadowing studies are consistent with language users' ability to make use of differences in the training stimuli and later show an effect in the direction of the training.

4.1.4. Categorical shifts are observed after exposure to higher-order linguistic stimuli.

Perceptual category plasticity has also been reported as on-line adaptations to rate as well as higher order linguistic structure. Although not directly addressed in the current work,

[21] There were also significant effects of presentation order found such that the most recently presented tokens were judged as more similar to the training stimuli than tokens presented earlier.

speaking rate has been shown to influence listeners' perception of linguistic categories, indicating again the pervasiveness of plasticity in categorical speech perception. Summerfield (1981) reported that as rate decreases the VOT category boundary shifted to become longer. In later work, it was also shown by assessment of goodness judgments that the best exemplars of tokens in the voicing domain also shifts toward longer values as rate decreases (Miller & Volaitis, 1989; Volaitis & Miller, 1992), and that this effect is more pronounced in the voiceless series than in the voiced (Miller, Green, & Reeves, 1986). These rate-dependent effects, especially those shown to be more pronounced in the voiceless series than in the voiced, are discussed below as possible explanation for results in the present work in which there is greater sensitivity to the frequency specification in the voiced series where less variability might be expected. That is, the reports above generally confirm the plasticity of linguistic categories, especially VOT, and, at the same time, lay the groundwork for the discussion below further explaining the frequency asymmetries found in the present work.

Another, somewhat higher-order linguistic domain of plasticity is observed in the phenomenon known as domain-initial strengthening in which phones at an edge of discourse-level organizational units of speech (Cho & Keating, 2001) are observed to behave differently than the same or comparable phones medially (Fougeron & Keating, 1997). For example, Fougeron and Keating (1997) report that a phrase-initial nasal in the same repeated syllable is strengthened (longer, for example) at the beginning of an utterance versus in the middle of an utterance. Cho and Jun point out that strengthening provides "perceptual cues for the prosodic structure and information groupings...[and] seems to help listeners segment the incoming flow of speech into smaller units, and recover the meaning of the utterance and the speaker's intention. Such an enhancement strategy may also facilitate lexical access to the domain-initial item which generally has less contextual or discourse information" (2000: 41). Taken at face value, if production differences are at least in part for the benefit of adding perceptual information, listeners are likely to be able to recognize these changes—including increased VOTs (Cho & Keating, 2001)—and take advantage of the information they provide. Furthermore, since the listener has no control over the production mechanism, variation due to strengthening in a temporal feature such as VOT must be dealt with rapidly on-line. Thus, for the present discussion, this on-line perceptual ability must be able to dynamically recognize a strengthened feature (such as lengthened VOT) as a category member despite a potentially unusual form, and the perceptual system must be able to cope with this variation. This contributes to the present results in that it helps explain how listeners are able to home in on specific auditory cues in the signal (the lengthened VOT, in this case), and encode those cues into linguistic memory. That is, the results reported here are not likely to be the result of factors not considered here, but are likely the direct result of the specific training used in the present experiments. Under an exemplar model, those encoded cues in memory serve to update the exemplar cloud for that token, reinforcing its representation—which is precisely what was found in the current work on lengthened VOTs.

4.1.5. Categories shift as a result of accommodation within a discourse.

In a slightly larger linguistic domain, talkers and listeners have been shown to accommodate their speech to each other, indicating again the plasticity of categories used in speech. Perhaps the most obvious instance of accommodation is the well-known phenomenon of intentional 'clear speech' in which talkers adjust productions in an effort to increase intelligibility under conditions perceived to be less than ideal. Under clear speech conditions, talkers have been shown to produce a variety of adjustments including increased amplitude, decreased rate, and increased VOT (see Smiljanic & Bradlow, 2005a, for a brief summary of these and other reported effects of clear speech). From the perceptual side, it has also been shown that clear speech generally improves speech recognition, increasing intelligibility under a variety of conditions (Picheny, Durlach, & Braida, 1985). Specifically relevant to the present work, English clear speech has been reported to have increased VOT as compared to conversational speech not intended as clear speech (Smiljanic & Bradlow, 2005b). This is another piece of evidence that representation of linguistic categories are not fixed, invariant codes in speakers' memory for words.

Other accommodation studies include work in which words that are repeated in a discourse are shorter than control words by up to 23ms (Fowler et al, 1997). Reductions were observed in words that were repeated within an utterance but reduction was blocked by 'episode boundaries'[22] which are argued as serving to essentially reset the available reference frames. When these episode boundaries were reset, reductions were not observed (Fowler et al, 1997). What is particularly interesting about this work is that talkers produced a form that was reduced (including, presumably the VOT of that form) as a way to *add* overall information to the signal, while at the same time listeners had to be sensitive to the same adjustments, including different boundary locations along the

[22] 'Episodes' are "defined by the occurrence of such scene markers as "so the next main thing that you see"" (Fowler et al, 1997: 26).

voicing continuum, in order to achieve a comparable level of recognition. The unexpected link between acoustic reduction and an information increase is an example of the wide range of variation talkers and listeners must deal with for normal language use.

In another study, Pardo (2006) found that language users interacting with one another tended to converge on forms judged more similar to each other after repeated usage than the same words before interaction. Subjects performed a map task giving each other instructions from similar maps and repeating the same lexical items several times over the course of a given interaction. Another group of subjects found later productions more similar to each other than earlier productions were similar to any other productions. Among other factors influencing production variation such as talker sex and role of talker reported by Pardo, convergence is further evidence of production adjustment and variation in the structure of the categories used for speech recognition. Pardo's results are also consistent with an exemplar memory approach in which recent experiences serve to update and strengthen representations.

Finally, in another study showing categorical structure effects based on experience, Sancier & Fowler (1997) found that a Portuguese-English bilingual varied her productions in the direction of most recent language experience. After several months in Brazil, an observed talker's VOTs in both languages were more like Portuguese (shorter), but after an extended stay in the United States, VOTs in both languages were more like English (longer). Again, this is evidence that linguistic categories shift as a direct result of exposure to variation, and that these shifts are evident both in productions and perceptions.

Although there are certainly complexities in the interactions between talkers and listeners with respect to accommodation shifts in linguistic categories, the talkers in the preceding studies produce speech that is influenced by recently heard speech, such as more immediately local speech with longer or shorter VOTs.

4.1.6. Categories shift as a result of exposure in the perceptual learning paradigm.

Finally, in a paradigm very similar to the present work, the plasticity of category structure is demonstrated in terms of perceptual learning. Generally, the small but rapidly growing perceptual learning literature trains[23] a naive listener to a relatively small number of exemplars that contain a manipulated target feature. Post-exposure tests of category boundary location reveal categorical shifts by comparing pretraining baseline values with posttraining values. During the training phase, perceptual learning uses perceptually ambiguous exemplars drawn from the boundary region of an auditory continuum indicating a single, binary perceptual contrast. Similar to the methodology used by Ganong (1980), the ambiguous presentation is lexically biased such that identification into one category results in a lexical item and categorization into the other results in a non-word (although both forms are presumably phonotactically licit forms). The pre- and post-exposure tests normally consist of a lexically unbiased continuum, such as [ɛf−ɛs], the names of the letters 'F' and 'S' (Norris et al, 2003).

In the first, seminal study on perceptual learning for speech processes (Norris et al, 2003; previous literature addressed perceptual learning in the visual domain, see

[23] The perceptual learning literature is inconsistent when referring to the learning phase of experiments as *training*, *learning*, *exposure*, or *listening*. For the present work, I will refer to the learning phase of these experiments—including the present results—in terms of *training*. There is no implication, however, of feedback or consciousness to the goals of the tasks that is sometimes associated with the term.

Bedford, 1995), subjects performed a lexical decision task (the training) on a selection of about 20 synthetically created ambiguous [s]- or [f]-final Dutch words and non-words embedded in a list of about 200 total words presented steadily 2600ms apart. One training group heard an ambiguous sound in words in which [s] created an actual lexical item but [f] did not; another training group heard just the opposite, the ambiguous sound in which [f] formed an actual word but [s] did not. A control group heard the ambiguous sounds where neither [s] nor [f] formed an actual lexical item. In posttraining testing on a categorization continuum between [ɛf] and [ɛs] (the names of the letters 'F' and 'S' in Dutch) , subjects reported ambiguous tokens favoring the bias toward real lexical items from the set each was trained on, either [f] or [s], critically depending on the training. The control group showed no bias or significant effects. The results indicate that listeners actively tune their perception to match the productions they are exposed to, those tunings critically informed by the lexical status of the training items. One particularly remarkable fact about this study (and subsequent work in this paradigm) is the apparent rapidity subjects' perceptual systems were tuned to auditory changes made "in an idiosyncratic way"; that is, after only about 20 of these odd, ambiguous presentations in a somewhat unnatural task (i.e., phoneme monitoring), perceptions were shown to be affected in the direction of the lexical bias and not biased in the direction of the opposing non-word. It is important to notice that in order to identify posttraining items differently from pretraining test items, the category and associated boundary delimiting these items must undergo change, and, since that change indicates a shift, the plasticity of the category is demonstrated.

These results have been replicated in a number of other studies (Kraljic & Samuel, 2005, 2006, 2007; Eisner & McQueen, 2006; Maye, Aslin, & Tannenhaus, 2003), and are idealized Figure 12 below. The first panel in the figure shows the initial pre-exposure normal response curves for a binary-valued category (*F*) along an appropriate acoustic dimension. The middle panel shows the training phase in which tokens previously determined to be perceptually ambiguous between +*F* and –*F*, which, in this case, are lexically biased such that the +*F* category constitutes an actual word but the –*F* category does not (the direction of bias is arbitrary). For example, in one study, productions of the fricatives [s] and [ʃ] were blended in variable ratios to create a lexical pair such as *brochure-brosure* in which one end of the continuum, *brochure,* is an actual lexical item and the other, *brosure,* is not (Kraljic & Samuel 2005). In the figure below, suppose for this example pair that the non-word s-category is represented by the –*F*, and the actual lexical item ʃ-category by +*F*. In the training phase, an auditorily ambiguous item between the two categories is presented that is lexically biased toward +*F* / *brochure*. The bottom panel shows the pre- and posttraining boundaries and the indicated leftward shift of the posttraining boundary toward +*F* / *brochure*. That is, after training on the lexically-biased but auditorily ambiguous tokens, listeners expand, move, or strengthen their response function to that lexical category.

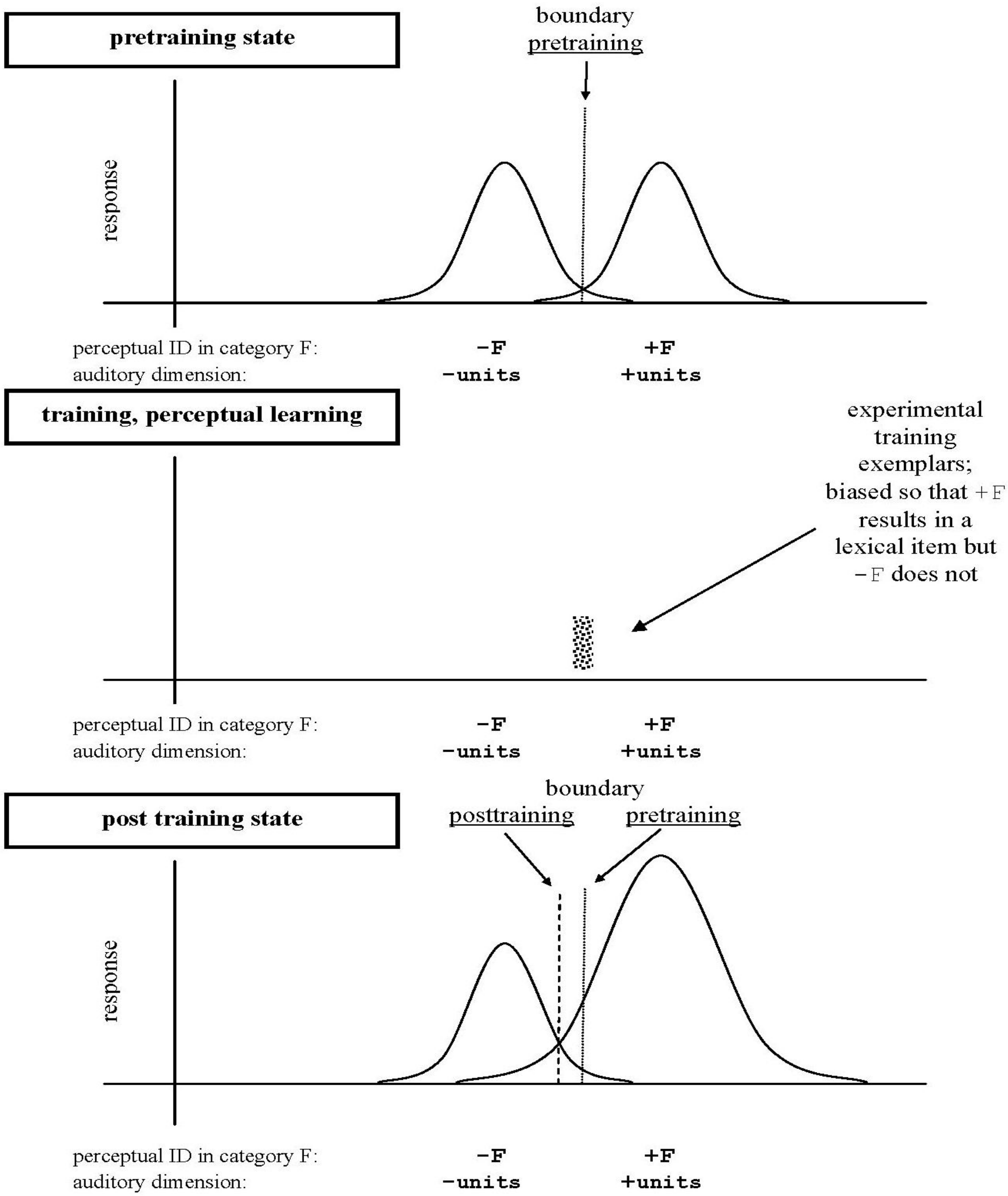

Figure 12. Schematic response curve shift after perceptual learning. The top panel shows the hypothesized baseline state for two categories defined along an auditory dimension (in units) and a binary perceptual category labeled 'F'. The middle panel shows the training phase in which an auditorily ambiguous, but perceptually lexical-biased token is presented (usually a small number of times). The bottom panel shows the resulting response curves after training, in which the trained-on category response function is increased (larger). The increased response is interpreted as a (leftward) boundary shift toward the relatively smaller category, indicating the trained-on category is larger in post-training testing than in pre-training testing.

A number of studies were undertaken extending the work begun by Norris et al (2003) showing in greater detail how "listeners adjust their phonemic representations to reflect the speech that they are exposed to" (Kraljic & Samuel, 2007: 1). In all these studies, listeners are shown to exhibit a presumed lexical bias[24] in the direction of training. Similar to the perceptual studies reported above, the present study shows categorical biases as a result of training on specific tokens. However, the results of the current work are the result of auditory bias to lengthened training tokens, rather than assumed lexical bias. That is, in the present study, the bias from auditory stimuli (and, by extension, also the lexical status of the target word) is more direct.

4.1.7. Summary of categorical plasticity

After framing the notion of categorical plasticity within the appropriate temporal frame, between hours and days, the preceding section discussed a number of ways in which linguistic categories have been shown to be flexible in the perceptual system of language users. Taken together, research bearing on category plasticity suggests that the phenomenon is robust and well-attested. The goal of this review is to demonstrate the wide variety of ways in which categorical perception is affected by experience and to provide a basis for understanding the results of the present work in terms of category plasticity. The design of the present work allows for subjects to be trained on auditory stimuli without a presumption of lexical bias. Results indicate that subjects are

[24] It is worth noting that in all perceptual learning studies, there is no attempt other than the implications of the study that listeners actually do access their lexicon. The possibility remains that a target word was unknown to a listener, or that a presumed non-word carried semantic content for some subject.

influenced by training on lengthened tokens, in the expected direction of training, offering evidence especially in accord with results from perceptual learning. Finally, these results are consistent with exemplar models of linguistic memory. As subjects are trained on lengthened tokens, their representations of those categories are strengthened and the categories expand, reflecting this experience.

4.2. Categorical plasticity effects considered in the present study.

Consistent with the work described above, the results from the present work show that the categories used in spoken word recognition are subject to experience. After training on lengthened target words in the present work, the boundary location for those words was shown to shift a small amount, about 3.3ms longer than before exposure. This primary result is important for three reasons: category plasticity, rate of change, and lexically specific effects. Each of these primary results will be described in this section.

4.2.1. Category plasticity is observed after training.

As described above, this primary result of the present work is that subjects' voicing category boundary shifted in response to specific training. This effect can be accounted for by an episodic memory model in which categories consist of clouds of exemplars sharing some feature or features accumulating from experience (or training, in this case) and losing representational strength as the result of decay. As a relatively small number of additional exemplars are added to the representational cloud, the addition of the new exemplars should contribute a relatively small influence on the exemplars in any given category. Thus, the observed overall category shift for those subjects exposed to lengthened exemplars in the direction of the exposure is observed. Presumably, exposure to the new, lengthened tokens influenced the representation of the category in the direction of the exposure.

Interestingly, a similar effect was not obtained for those subjects trained on synthetically shortened target words. If memory was affected by training on lengthened exemplars, why do we not also find a shortened effect for those trained on synthetically shortened exemplars? This lack of effect, however, is likely due to a number factors not inconsistent with exemplar memory. First, shortening is different from lengthening with respect to how a contrast is auditorily defined. Shortening the VOT of a voiced speech sound makes that sound auditorily more similar to the contrasting voicing sound. On the other hand, lengthening VOT of a voiced item makes that item more auditorily dissimilar from the contrasting voiced sound. Furthermore, there is an absolute limit to the degree voiceless VOT can be reduced (both in terms of the relationship to the contrasting sound's acoustic values and relationship to a physical floor of zero-VOT). There is no concomitant principled ceiling with respect to the degree voiceless VOT can be lengthened. Second, the percentage of shortening was considerably less than the percentage of lengthening. The overall mean natural VOT of target items in the training story is 73.5ms. The lengthened training words average 132.2ms (+80%), and the shortened words average 58.8ms (−20%). The degree of acoustic reduction, 14.7ms versus 58.7ms, is likely relevant to the perceptual system. Given the modest effects of the lengthened training condition, a more similarly scaled effect might be expected in the shortened condition. Whether those effects are scaled by percentage reduction (80% versus 20%) or by average training-to-effect-ratio (3.3ms change for 58.7ms exposure) is unclear, but likely lost to values below the perceptual or statistical threshold. Finally, recall that no subject was ever trained on the naturally-produced training story, so comparisons in terms of reduction is inappropriate since subjects never heard the original recordings. Rather, the shortened condition might best be interpreted as a near-natural control condition. In any case, the lack of effect in the shortened group is unproblematic to the positive results of the lengthened group.

4.2.2. Rate of category shift is fast: only a small number of exposures is required for category shifts.

Successful training is evidence that a small number of training exposures is required for category plasticity. The experiments reported here trained naive listeners[25] to only about 40 total training exposures[26] to the artificially lengthened exemplars. This scant exposure coupled with robust effects of training in the lengthened group is strong evidence that category plasticity in spoken word recognition is rapid. Opposed to a traditional view of stubbornly static, absolutely invariant categories (or features), this result is compatible with an exemplar memory for speech that is constantly updating representations for words. Similar recent research in the perceptual learning paradigm developed by Norris et al (2003) (Clarke & Luce, 2005; Nielsen, 2005, 2006) have shown similarly rapid acquisition of plasticity in spoken word recognition. Finally, although exemplar theories posit a decay function in which unused representations are weakened over time, the

[25] In informal post-experimental interviews, not a single subject reported conscious awareness of the goals of the experiment or awareness of synthetic tokens in the training. On several occasions, I described the lengthened VOT in target words, replayed segments containing the target words, and subjects refused to believe the signal to be modified.

[26] As mentioned earlier, it is possible that the reported effects could be the result not of the accumulated effects of the duration of the experiment, but the result of perhaps the last day only. This possibility, if verified, would be stronger evidence of the claim being made here. In either case, however, the rate of training must be considered very fast.

present study did not address this issue. Further research would likely benefit from replication of the present work including follow up work investigating the rate of decay, or, in this case, how quickly a shifted boundary returns to its pretraining position.

4.2.3. Successfully trained subjects do *not* generalize training to new lexical items.

The present work suggests that training affects specific lexical items and does not generalize to other, similar lexical items, or, more generally, that those effects do not propagate throughout the lexicon as predicted by the traditional model. A similar lack of generalization is reported by Clarke & Luce (2005), but a different conclusion is reached by Nielsen (2005, 2006) and Kraljic & Samuel (2006) working in a similar paradigm. Subjects in the present study were given two post-training voicing identification tasks, one on the same target lexical items used in training, and another on a similarly matched set. For those subjects who were successfully trained (those trained on lengthened exemplars), a similar boundary shift in the new words would be evidence that training was generalized to the new words, and, possibly, throughout the lexicon. If, on the other hand, boundary identification for words in the newly presented set do not show a similar boundary shift, the effects of training must be understood to affect specific lexical items without regard to similar lexical items.

This logical test juxtaposing generalizability and lexical-specificity, however, is complicated by the need to strictly limit exposure to the exact items used for testing after training. If both word sets (the set used in training and the matched set) are tested before training to establish baseline values, subjects encounter both sets of lexical items and absolute frequency, attention, and familiarity are potentially affected. Thus, to eliminate this possible exposure effect during pretraining, boundary values for words used in the posttraining generalizability portion of the experiment were not collected. Testing for effects with no baseline for comparison, however, is easily accomplished by comparing between training groups rather than within training groups. This method has the advantage of allowing for relative comparison between experimentally novel forms which further allows for broader possible differences between the two lexical sets (such as a non-word in one set actually occurring as a slang or dialectal form for some subjects). The disadvantage of this method is that absolute values cannot be determined, since only relative comparisons are anticipated. Thus, the test simply looks at the relative boundary values for the two sets *between* the two training groups. If the boundary values of the posttraining, novel word set are similar between training groups, the training effect is lexically specific; if the boundary values of the posttraining, novel word set are dissimilar between training groups, the training effect has been generalized. Of course, if a generalization effect is found, it must be checked for direction either toward or away from training.

Results indicate that the boundary values of the posttraining, novel word set are similar between training groups, so the training effect is lexically specific and not generalized. These results are schematized in Figure 13 below summarizing the mean VOT boundary values among training groups (shortened and lengthened), training session (pretraining and posttraining), and word set (training set and generalization set).

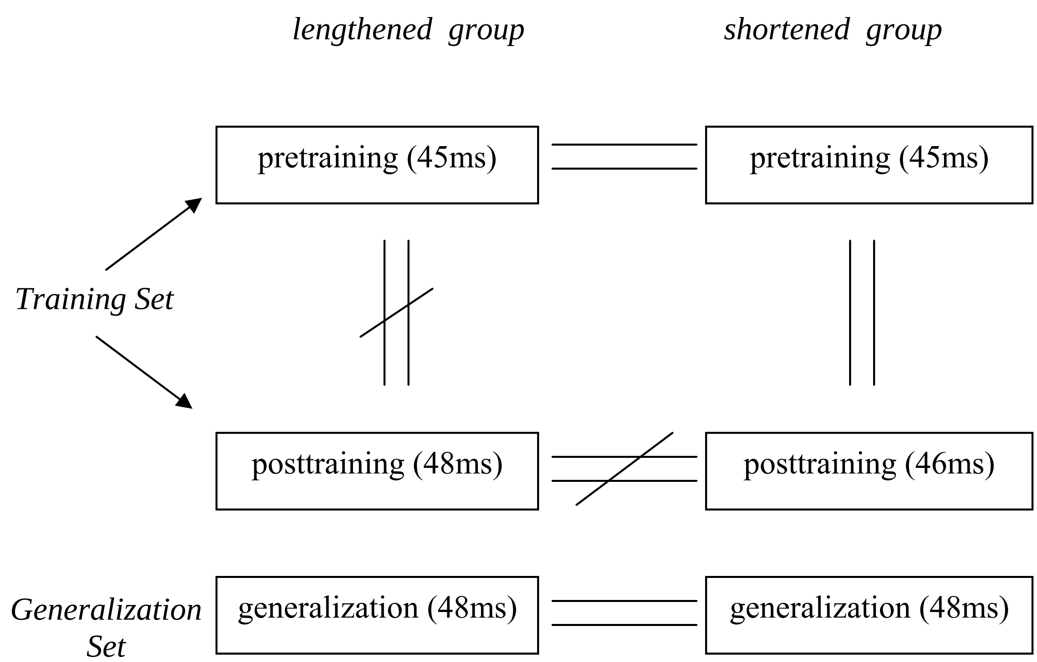

Figure 13. Effects by group (shortened, lengthened), lexical set (training, generalization), and session (pretraining, posttraining, generalization). Parallel lines between boxes indicate no significant difference between groups, crossed parallel lines indicate a significant difference between, and no line indicates no possible relationship. Overall, the lack of difference between the two groups in the generalization condition indicates that subjects successfully trained on lengthened exposures did not generalize that effect to (similar) new lexical items.

Several features of Figure 13 are important. First, the posttraining effect of 3.3ms for the lengthened group is shown in the inequality between pretraining and posttraining for the lengthened group, and the same effect is not observed in the group trained on shortened target words (and, although not significantly different, is in the unexpected direction given previous results). Second, the posttraining condition for the lengthened group is dissimilar to every other group using the same training set of words. The posttraining condition for the lengthened group is the one value not like the other values in the four training set testings. Third, the pretraining values for both groups of subjects are similar, and directly comparable since the same set of training words is used for all subjects. So, the lexical items in the training set are shown to be unbiased for either group in pretraining testing, but biased for the lengthened group after training. Similarly, when the experimentally novel generalization set is introduced after training, the between-group boundary values are directly comparable, but, since the same set of words is not used between the training and generalization sets, a within-group comparison (of absolute values) is not possible. Since the generalization conditions are in fact nearly identical (as they are in the pretraining condition), it is concluded that subjects do not generalize the effects of training to new lexical items.

The lack of generalization effect is consistent with an exemplar model, but difficult to explain in the traditional, abstractionist model since any distinction is spelled from invariant phonological segments and distinctive features. Since only the lexical items that were trained on show the training effect, an exemplar model predicts that training should targets only those particular items. Also consistent with an exemplar model, structurally similar lexical items, the words in the generalization condition, have unbiased representations and do not undergo change. Presumably, the training reinforced the representational strength of some specific lexical items, but did not propagate more broadly throughout the lexicon. In the traditional model, if a change is evidenced, as it is for those in the lengthened condition, that change should affect every word with that specification, including the words in the generalization condition. Since there is no lexical propagation, the traditional model is unable to explain the data.

4.2.4. Importance of training effects.

Before the experiments were conducted, it was unknown if a small number of training exposures to naive subjects could affect their representation of speech categories using the methods of the present work. These experiments confirm other data suggesting that language users' categories are indeed plastic and undergo change in response to experiences. Furthermore, the results suggest that the changes are lexically specific, targeting specific lexical items but not others. These results are compatible with the exemplar model of memory for language: as new episodes are experienced, certain features of the exemplars are added to (or update) the representation of language in memory.

4.3. Frequency effects of the present work considered.

Another important variable in the present work is lexical frequency. Lexical items used in the present work are paired with a rhyming word contrasting in the voicing of the initial segment. Each word in each pair has an associated lexical frequency: high, low, and zero frequency non-words ('no-Frq') for the voiced words, high and low for the voiceless words.[27] This organization is given below in Table 6 with representative words used in the experiments.

[27] Recall that the reason for this design asymmetry is that the distribution of voiceless stops in English is greater than for the voiced stops. That is, there are more actual lexical items with voiceless stops than with voiced stops. Because of this lexical asymmetry, fully specified sets of matched test items were not possible to construct.

ALVEOLAR	no-Frq [d]	lo-Frq [d]	hi-Frq [d]
lo-Frq [t]	*teak ~ deak*	*taupe ~ dope*	*ton ~ done*
hi-Frq [t]	*top ~ dop*	*ten ~ den*	*too ~ do*

Table 6. The organization of word pairs by lexical frequency.

To test for effects of frequency, the identification task yielded category boundary values which could be used to compare words of differing frequency. Since contrasts are crucial for the identification tasks, the (voicing) contrasts are necessarily defined by two actual presentations. As a result of this design, lexical frequency is examined from a number of perspectives described below. First, the importance of the frequency specification of the voiced member of the pair, but not the voiceless is discussed with possible reasons for this asymmetry. Next, the Ganong effect is discussed with respect to both lexical bias and frequency bias. Results from the present work do not replicate the lexically biased Ganong effect, but, in an apparently straightforward extension of previously reported Ganong effects, a frequency bias is described.

<u>4.3.1. There is an asymmetry in the perception of voiced and voiceless frequency.</u> Results indicate that the frequency specification of the voiced member of a contrasting pair might contribute to perceptual identification such as in the tasks reported here, but the frequency of the voiced member of the pair might not be used in the same way. This result is displayed in Figure 14 below, replicating the portions of figures given in the results section. This figure shows the VOT boundary location on the ordinate and frequency specification by voicing on the abscissa. Frequency specification of the voiced member of a pair is shown on the left side of the panel, and the frequency specification of

the voiceless member on the right. Non-words ('no-frq' words are not included in the display).

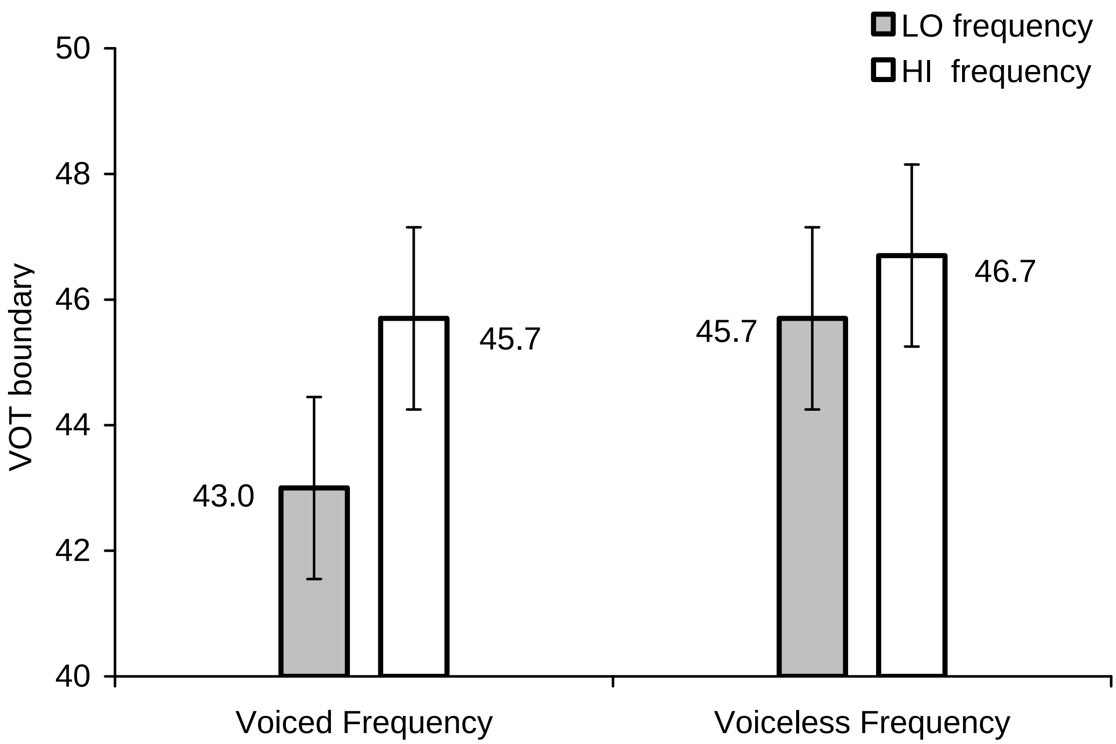

Figure 14. Lexical frequency effects of words by voicing. There is a significant difference between the frequencies for the voiced member of a word pair, and there is not significant difference for the voiceless frequency specification. Notice, too, that in addition to the voiceless frequency lacking significant difference, the direction of observed difference is in the wrong direction.

The figure shows a significant main effect of frequency (2.7ms) for the voiced member of the pair, but no significant effect for the voiceless member. The effect in the voiced frequency specification is also in the expected direction, indicated by a longer VOT boundary for high-frequency voiced words as compared to low-frequency voiced words. The longer boundary value indicates that more high-frequency words were labeled voiced than low-frequency words. The lack of effect for the voiceless frequency specification (and in the wrong direction[28]) indicates no effect of boundary difference when subjects are presented with differing frequencies of voiceless words.

Although this asymmetry was not expected, there is no apriori reason such an asymmetry should not exist, and evidence from the literature offers a description of a similar asymmetry that might help explain the present results. In a production study on syllable timing and VOT, Miller et al (1986) explored the relationship between syllable length and voicing category structure. They report that as subjects produce longer syllables, the VOT value for the voiceless member of a pair increases, but the VOT for the voiced member remains constant. In their study, as overall syllable duration increased, the voiceless category shifted dramatically longer, but the voiced category remained remarkably unchanged. The relevant display from their study is shown below in Figure 15. In the display, the triangles represent the voiceless tokens and the lower asterisks represent the voiced tokens. Notice that as syllable duration increases, the VOT for the voiceless series has a noticeable positive slope, while for the voiced series there is no slope. Notice, too, that the error bars reported by Miller et al (1 SD +/- mean) are larger at longer rates, indicating more variability as syllables and voiceless VOT becomes longer.

[28] When comparing two boundary location values, a longer value indicates more 'voiced' responses, and shorter value indicates more 'voiceless' responses. Thus, if there were a frequency effect for the voiceless frequencies, the expected direction of the relationship is to a longer value for the low frequency specification, while for a frequency effect for the voiced frequencies, the expected shorter low-frequency value is observed.

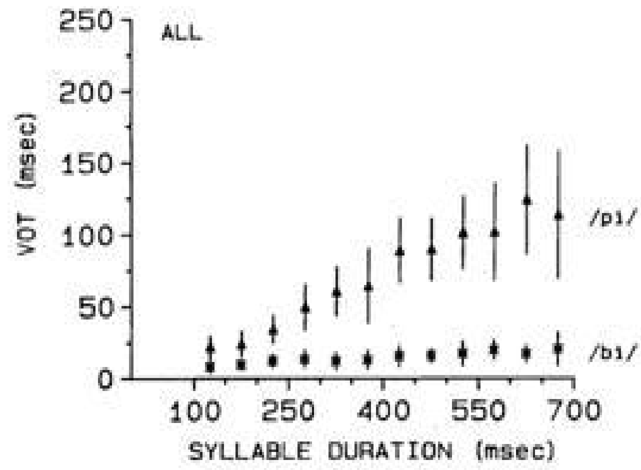

Figure 15. from Miller, Green, and Reeves (1986). The upper series represents production values from VOTs as longer syllables were produced, and the lower series shows the voiceless values along the same scale. As syllable length increases, only the VOTs of the voiceless words also lengthens, while the same syllable lengthening does not increase the VOT duration of the voiced series.

One important result of this study is that it shows greater flexibility in VOT values for the voiceless member of a pair, both overall and in rate-dependent factors, but, when longer, there is also apparently more variability. Based on these results, listeners are likely to expect less variation on the voicing end of the continuum, and might be more sensitive to idiosyncratic variation there since less might be expected. Furthermore, since any variation in the voiceless series is apparently not tied to syllable duration, detected variation can more easily be utilized to indicate other information such as lexical frequency. By contrast, if duration of voiceless VOT values is tied with overall syllable duration, listeners would be less likely to be able to decide whether a perceived lengthened VOT is due to reduced rate, lexical frequency marking, or other features.

With respect to the present work, then, sensitivity to variation in voiced words might be expected, but variation in voiceless words might be less likely. That is, although not directly tested in the present work, subjects might be more easily trained on

VOT manipulations in the voiced category as opposed to the voiceless category. Although this presents an asymmetry in the system, there is no apriori reason to prefer symmetry. Furthermore, in an exemplar model, there is no inherent or necessary link between two features that vary along an auditory scale such as VOT. The possibility that some words are more sensitive to changes is not precluded by an exemplar model. On the other hand, the traditional model has great difficulty dealing with any lexically specific specifications such as these reported here. Given this asymmetry, the next issue is to look closely into the frequency differences within the voiced series.

4.3.2. Frequency differences in the voiced series.

As mentioned above, there is a three-way mutual contrast among the three lexical frequency specifications in the voiced series, but no difference between the specifications in the voiceless series. Since the variation is restricted to the voiced series, the following discussion will be restricted to the voiced series only.

There are two important observations regarding details of frequency effects in the voiced series, both relating to the Ganong effect. The Ganong effect has been reported when an auditorily ambiguous token is shown to have a lexical bias on a word- versus non-word continuum (Ganong 1980). In a straightforward extension of the Ganong effect, we found that auditorily ambiguous words have a high-frequency bias on a high- versus low-frequency continuum, as described in the previous section. We further expected to replicate the lexical bias of the Ganong effect when comparing words versus non-words. Surprisingly, the Ganong effect was not replicated, and a possible anti-Ganong effect is described. Each of these effects is described in more detail below.

Closely related to earlier work on phoneme restoration (Warren, 1970), the Ganong effect makes the claim that given the choice between word and non-word, people tend to favor the word. The issue in the literature generally focuses on whether this effect and related results bear on the nature of lexical access and linguistic processing (Ganong, 1980; Samuel, 1986; Elman & McClelland, 1988; Pitt & Samuel, 1993; Norris et al, 2003; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006, 2007). Although previous research has been conducted under a variety of conditions and for a variety of goals, the basic observation of lexical bias appears to hold up. A critical figure from Ganong's (1980) paper is given below, nicely illustrating the effect. In Figure 16 below, the dashed line represents the boundary between a nonword-word pair (such as *beace-peace*) and the dotted line represents a word-nonword pair (such as *beef-peef*).[29] Each response curve is shifted in favor of the actual lexical item, making the category of the lexical item larger. Since the shift affects only the acoustically ambiguous tokens (the tokens on the endpoints asymptote and are not relevant), an ambiguous token is more likely to be categorized as the lexical item than as the non-word.

[29] Ganong reports that "the solid line shows idealized data from a neutral continuum, whose perception is not biased by lexical effects....In the present experiment, the phonetic categorization of a lexically biased continuum is not compared with the categorization of an unbiased continuum, but with a continuum biased in the opposite direction" (1980: 112). Notice that the present experiments are similarly designed such that evaluations are made within- and between-group, but not with respect to a 'neutral' baseline.

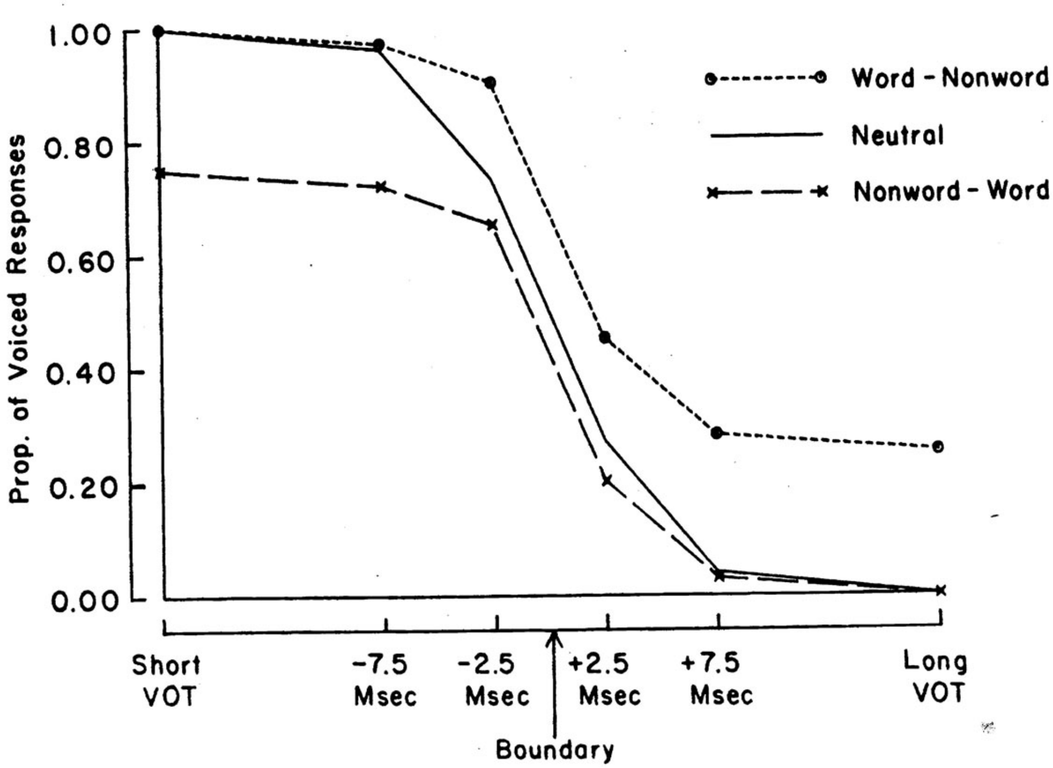

Figure 16. from Ganong (1980). The relative locations of the dashed and dotted identification functions indicate relatively shifted category structure, including estimated category boundary location, based on the lexical status of a presentation form (word versus non-word).

Based on results of lexical bias and Ganong effects, the present work investigates the possibility of a similar bias based on lexical frequency of occurrence. That is, Ganong's experiments used pairs that differed such that one item was a word and the other was not. The present work extends this by using pairs that differ such that one word is a high frequency word and the other is a low frequency word. It was predicted that there would be a similar high-frequency-over-low-frequency bias. Results in the present work indicate that there was indeed a frequency effect, about 2.7ms favoring high-frequency over low frequency—although only for the voiced words, as described elsewhere. Notice that the effect is in the expected direction, favoring high-frequency words. That is, given an auditorily ambiguous token, listeners are more likely to categorize that token into the voiced high-frequency category rather than the voiced low-frequency category.

This bias favoring more frequent words, however, must be qualified. First, the frequency bias is only observed for the specification of the voiced element of a contrasting pair. As described above, main effects of frequency were not found when the more frequent item began with a voiceless stop. This lack of effect for the voiceless series is perhaps the result of the variability discussed above with respect to the larger variation in the voiceless series. Second, frequency effects have long been associated with auditory reductions (Schuchardt, 1885), but expected frequency effects preferring higher frequency words over lower frequency words have been shown to be smaller than expected (Cohn et al, 2005), or in some cases undetectable (Guion, 1995; Cohn et al, 2005). Although a significant effect was indeed found, the observed difference is a modest 2.7ms. Third, although it is not unreasonable to explore a possible relationship between the Ganong effect and a frequency effect, there is no reason to believe they are related to one another. The claim here is that the lexicon contains words arranged in some way according to lexical frequency; there is no claim that non-words are stored in the lexicon, however.[30] Since words with a lexical representation must have some frequency specification, and words not in the lexicon cannot have a frequency specification, it is unknown how recognition of these two different types of forms might

---

[30] A non-word item stored in the lexicon is not as ridiculous as it might at first sound. Since lexical representations (under all theories considered here) critically include both a form and a meaning, it is possible that a form is stored in the lexicon without any associated meaning. In that case, it would be possible to appeal to a specified form that is in fact a non-word (simply has no meaning). Also, people regularly learn new words and admit those new items to their lexicons. These issues are somewhat outside the scope of the present discussion.

differ in recognition. So, although on the surface a frequency effect might appear to be expected from results associated with Ganong effects, the relationship is tenuous at best, and possibly describe unrelated systems or phenomena. As will be discussed presently, different phenomena suggest different representational systems.

4.3.3. Auditorily ambiguous non-word bias: a possible anti-Ganong effect.

Recall the Ganong effect predicts a lexical bias in which an auditorily ambiguous token is perceptually biased to be heard as the word over the non-word in a contrasting pair. Due to restrictions imposed by the experimental design, non-words are only represented in the data sets as opposing elements in the voiced series. That is, word pairs such as *dake-take* are represented in the experimental data, but word pairs in which the non-word is voiced and an actual lexical item is voiceless (Ganong's *beef-peef* contrast, for example) are not tested in the present experiments. This design restriction, however, is not problematic in principle, since a lexical effect can still be tested.

In the present study, the Ganong effect can be directly tested with three possible analytic outcomes. First, replication of the Ganong effect would be indicated by a VOT boundary for non-words shorter than actual words (the combination of low and high frequency words, or either category independently). Another possible test result is that no significant effect between words and non-words is found. A non-effect would indicate non-replication of the Ganong effect, but would leave interpretation elusive due to the lack of results. Finally, an effect of non-word bias would be obtained if the boundary value for non-words is longer than matched lexical items. This final possibility, indicating significant difference preferring non-words, would be direct evidence against the Ganong-effect, a possible anti-Ganong effect. This last possibility is, to my knowledge, unattested in the literature.

As shown in Figure 17 below, the VOT boundary for non-words ('no frequency') words is *longer* than either the low or high frequency words, indicating a possible anti-Ganong effect. This figure shows only the voiced series by frequency and lexical specification. Non-words are indicated by the specification 'no frequency' and high and low frequency words are specified as above. When the voiced member of a pair is a non-word, the boundary is 5.5ms longer than the actually occurring lexical items (6.8ms longer than a low frequency word, and 4.1ms longer than a high frequency word).

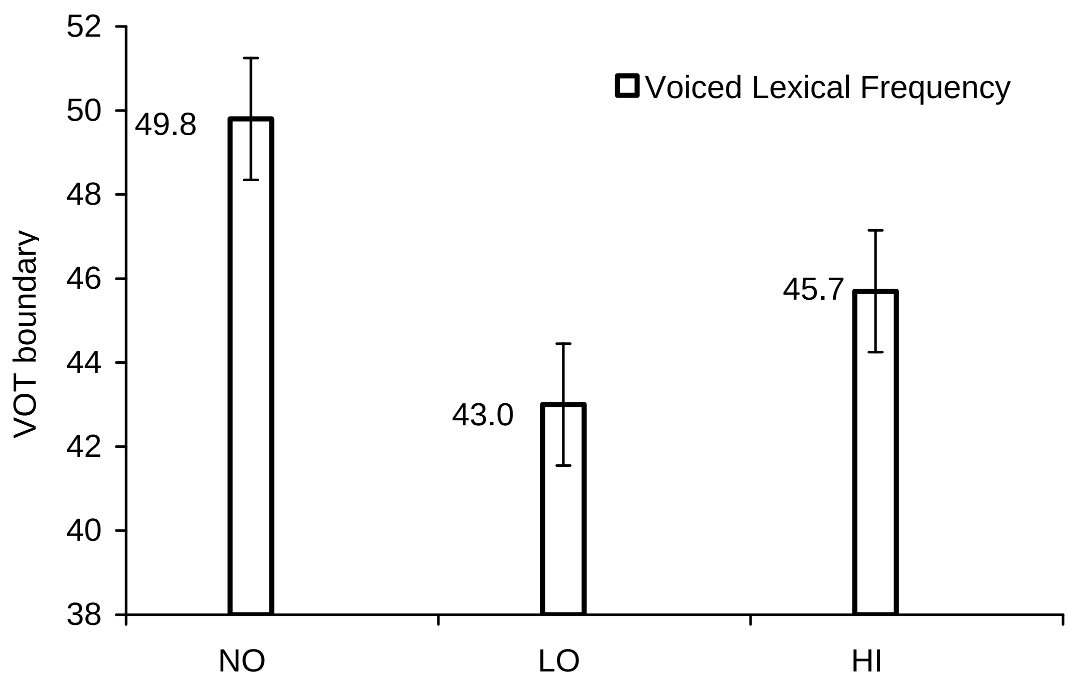

Figure 17. Lexical effects on VOT boundary for voiced words. Each of the three boundary values are significantly different from the other two.

This unanticipated anti-Ganong effect could be the result of a number of factors independently or in combination. One possible explanation is a novelty effect: subjects may have been responding to the novelty of non-words in the context of the laboratory and experimental setting. The identification task required subjects to perform a two-alternative, forced-choice identification. When presented with a non-word and a word (recall subjects were never presented with two non-words because the design precludes voiceless non-words), subjects may be reflecting a response bias due to special attention paid to novel items. Another possible source of anti-Ganong response bias might be the asymmetrical design of the experiment in which no voiceless non-words were presented. It is possible that at some level subjects noticed novel non-words only in the voiced condition and responded unpredictably to that asymmetry.

Although subjects were not made explicitly aware of any experimental design or asymmetry,[31] the assumption of experimental symmetry may have resulted in the observed bias. Another possible bias might be entailed in the set of items tested. Lexical frequencies and lexical status were collected from published sources (*The Hoosier Mental Lexicon*), but individual experiences with words are impossible to control for. For example, an experimental non-word such as *dake* could potentially be an actual (dialectal, slang, regional, etc.) lexical item for some subject or group of subjects. No posttesting data to check this possibility was collected. A fourth possibility is that the type of stimuli used in the experiments is unlike stimuli used in previously published work. In all previously published reports, stimuli employed for testing were synthetic and of rather low quality. In Ganong's seminal study, he reports using the Haskins

[31] Subjects were given minimal instruction, told only to identify each presentation as quickly and accurately as possible.

Laboratories speech synthesis by rule program, FOVE, created by Ingemann (1977) (Ganong, 1980: 114). This method of stimulus fabrication, common in the majority of psycholinguistic spoken word recognition research, results in highly unnatural stimuli. Observe Ganong's response curve reproduced above in Figure 16. It is evident from the figure that responses achieve substantially less than ceiling performance (less than about 80% for both synthetic stimulus curves). By contrast, the stimuli used in the present study are highly natural, fabricated exclusively from natural productions. Response biases observed by Ganong and others, then, is in response to a different kind of stimuli. A different, perhaps more ecologically plausible, anti-Ganong effect might be the result of such differences. Taken together, these factors alone or in combination might have influenced subjects response bias that yielded the apparent anti-Ganong effect.

4.4. Interaction effects among experimental factors: do frequency effects interact with training effects?

The discussion above considers a number issues with respect to both training effects and frequency effects. There remains the question of whether these factors interacted with one another in a meaningful way. In short, there were no meaningful interaction effects between training and frequency. Interpreting a negative result (the lack of significant interaction between training and frequency), however, is not a sound way to conduct this sort of research.

This non-result offers little opportunity for a specific interpretation in support or refutation of an exemplar or other model. Exemplar memory posits that representations for words are constantly updated based on experience. The strength of a representation is understood to be roughly the density function of experiences contributing to any representation: the more episodes of a token, the stronger the representation (Hintzmann, 1986). Thus, if a representation consists of a very large number of experiences or episodes, then its representational strength is high, and, critically, a small *absolute* number of exposures is unlikely to have an appreciable effect since the additional exemplars would contribute *relatively* little to the representation. On the other hand, a representation based on a small number of experiences is proposed to have an overall weaker representation, and a small number of new experiences should contribute relatively more to that representation. That is, the same absolute number of experiences should contribute less to the overall representational strength of dense exemplar clouds versus sparse exemplar clouds. Thus, it might be expected that contributing new experiences to a sparse exemplar cloud (that is, a low frequency word) would be more likely to result in that representation being affected. Since representational strength is directly tied to experiences, and frequency of occurrence is a measure of experience, higher frequency lexical items might be expected to constitute more stable, less malleable categories or words than low frequency lexical items (Goldinger, 2000).

Following this logic of representational strength and its relation to frequency, an experimental outcome of the present work that would have proved favorable to exemplar memory might show a training effect restricted to low-frequency lexical items. The observed effects, however, show effects statistically evenly distributed over both high and low frequency words. That is, the present results did not show a larger training effect on low frequency words than on high frequency words.

Importantly, despite the inability of the present work to offer support for a very detailed prediction of exemplar theory, the observed results do offer direct support for exemplar memory theories more generally: categories of target were indeed affected. Although the training results were broadly distributed among the targeted words, the lack of generalization effects described above demonstrate the ability to target specific memory for words.

Finally, it is worth mentioning that there are a number of factors not considered in the present analysis or discussion that are part of the experimental design. Such factors include stimulus place of articulation, longitudinal learning effects, response latency, and stimulus talker sex. To the degree that these factors have been investigated, no substantial or surprising effects have yet been discovered. Future research will look more closely at aspects of the data that appear to be promising. Currently, those areas of promise include the unanalyzed production data, response latencies, goodness judgments, and longitudinal responses.

4.5. The traditional model versus exemplar model.

Described in the introduction, the motivation for this work is to test two basic, critical assumptions of the traditional linguistic model: universal, invariant features and economy. The results of the present work offer evidence that neither basic assumption holds while at the same time offering evidence in favor of rich memory or exemplar model.

4.5.1. Plasticity of features can be accounted for in an exemplar model.

Under the assumption of a universal feature set, a proposed feature such the feature employed for the voicing contrast, is prohibited from undergoing any change in how it represents that contrast since that contrast is predefined. This is directly challenged by evidence in the current study that subjects were successfully trained on exposure to lengthened tokens in a specific set of words. Trained subjects show a small overall effect of category plasticity averaging 3.3ms in the direction of training. That is, subjects are shown to have dynamic category structure that shifted after exposure. Critically, subjects not trained on lengthened tokens did not undergo the same change, so the change is unlikely to be rooted in a factor other than experience. Furthermore, it is remarkable that the shift is observed in naive subjects after only about 40 exposures. Taken together, the present work offers strong evidence that the representations or language mechanisms responsible for contrast is not an invariant, universal feature.

These observations offer a serious criticism of the traditional model while at the same time offering support of exemplar models for language. In an exemplar model, representations are based on storage of detailed auditory features, and the strength of those representations crucially depend on experiences. New experiences update the category representation, and, as a consequence, directly affect the representation of those items in memory and these affect perceptual judgments. The present work shows that explicitly targeting specific representations can result in different identification response functions to the categories that rely on those representations.

4.5.2. Linguistic memory including richly specified features can be accounted for in an exemplar model.

Under the assumption of economy in the traditional model, a small set of features is assumed to represent all language contrasts. If a large number of detailed distinctions are shown to persist into the lexicon, economy is violated. This, too, is directly challenged by the present work in which a large number of contrasts are shown to be lexically- and frequency-restricted, and, after undergoing structural change, not generalized throughout the lexicon.

As described above, listeners are sensitive to the lexical and frequency properties of specific tokens. Lexical sensitivity is observed in the anti-Ganong effect and frequency specificity is shown by the different boundary values of high and low frequency words in all experimental conditions. Since that sensitivity cannot reasonably be attributed to phonological structure or generated by phonological rule (word sets differing by frequency and lexicality are matched in phonological structure), the observed variation must be represented in the lexicon, adding to the overall amount of lexical representation. Thus, the overall number of additional features must necessarily increase, critically violating economy.

Similarly, affecting the categorical structure of some targeted words did not generalize throughout the lexicon, also violating the notion of economy. If the lexicon is represented by a small number of features, changing one feature is expected to have consequences affecting all words containing that feature (and some structural description). For example, if the feature representing the voicing contrast is changed, then every lexical item containing that feature should be affected by that change. In other words, under economy, lexically specific changes (i.e., lexical diffusion) are not expected. Critically, if structurally matched words assumed to share a feature are affected differently, that difference cannot be represented in memory by the same feature. Those words treated differently must be differently represented, and, as a consequence, increase the overall number of features, critically violating economy. The present work offers evidence of just this kind: only words targeted in the experimental training conditions showed category shifts.

Similar to conclusions reached with respect to dynamic features, the clear violations of economy are evidence against the traditional account as described above, and simultaneously serve to support an episodic memory account. Episodic memory proposes lexical items are represented in memory by a large number of exemplars, which are, by extension, organized in memory using a potentially unlimited set of features. Thus, episodic accounts do not demand economy, so richly specified lexical memory as evidenced by the present work is consistent within the theory.

4.6. Summary of results and implications.

The present work shows that listeners' linguistic representation of words in memory includes factors such as lexical- and frequency-specific features. These features are shown to be highly specific, not generalized throughout the lexicon, and subject to specific linguistic experiences that affect their representation in memory. These results present a theoretical difficulty to the traditional, abstractionist language model but are consistent with episodic language models.

The results have a number of theoretical and practical implications for the study of language. First, this work has implications for theoretical modeling. There is a growing literature that memory for language, linguistic access, spoken word recognition, and speech production does not rely on abstract, invariant representations similar to orthography or transcription. The present work adds to this literature and suggests directions for improved language modeling and understanding of the underlying mechanisms of language including spoken word recognition, speech perception, and language processing. Second, this work has implications for first and second language acquisition, suggesting that the theoretical perspective critically tied to invariant features might benefit from a broader understanding of language. Further, both practical and theoretical language acquisition work might be able to benefit from the training procedures in the present work, offering new methodologies to explore areas of interest. Third, the present work offers potential insight into language change and variation. Results indicating rapidly changing and highly plastic representation of language is likely to influence how researchers could investigate rate and degree of linguistic change. Fourth, the present work has the potential to contribute to computational language models and applications, including speech related human-computer interaction. A better understanding of the processes underlying speech are undoubtedly an asset to this research.

5. Appendix

baby **talc** feels nice on the skin
shoe **talc** frequently dries the leather
creamy **dalc** should not be thrown out
three **dalc** sheets were formed by hand

a utility **tine** sharpener is easy to operate
a few **tine** shards broke off in the water
they **dine** silently with each other
Fred and Sally **dine** selectively on vegetables

the flimsy **teal** fern is the most prized
a fancy **teal** handkerchief can be silk
a mafia **deal** sometimes will lead to trouble
cobra **deal** swiftly with mice and rats

millennia **take** several generations
they **take** their time to get ready to go
fiery **dake** sends embers into the air
rosy **dake** seemed out of place there

the ordinary **time** varies according to the scale
ordinary **time** is kept by wristwatches and clocks
no **dime** fell into the reflecting pool
an extra **dime** vanished in the dark

no **town** seems to be safe from viruses
a sea **town** smells like the water
go **down** somewhere into the room
turn the radio **down** from the console

Table 7. Alveolar training production sentences.

free **kiln** space is difficult to find
a blue **kiln** heats up quickly
dirty **giln** scraps littered the floor
flashy **giln** supposedly looks like copper

they **caulk** shower stalls only in the mornings
the shiny **caulk** sparkles in the moonlight
silently **gawk** from behind the counter
you **gawk** so long they must know you're watching

move the yellow **cot** very carefully
no **cot** sleeps nine people
they **got** seven recommendations
anxiety **got** zapped quickly by his mood

they **keep** some money in the back room
cats frequently **keep** themselves clean
tomato **geep** smells funny
foamy **geep** varies quite a bit

they **call** from Canada every night
go **call** some of your friends
his pushy **gall** forced them to leave
his extra **gall** vexed the campers

Barry **could** shave in under 20 minutes
his tattoo **could** frequently be seen under his shirt
no **good** shells are inside the lagoon
my **good** friends are dependable

Table 8. Velar training production sentences.

the yellow **tint** follows a pattern
the runny **tint** seems too dark
algae **taint** their surroundings with color
panda **taint** the forest floor if left alone
Eskimo **tuque** feathers are unique
the new **tuque** visor is small
any **tot** should not be left alone in a car
the happy **tot** sitting on a table is laughing
they **took** five spoons from the jar
Jeremy **took** some time off work

go **cull** the rotten ones from the bunch
they **cull** some inferior fruit from the batch
salty **kale** smells like a wet vegetable
sea **kale** varies from place to place
a Lao **kip** value is less than a penny
the early **kip** scale was well known in Britain
the **keel** should always be above the waterline
jumbo **keel** sections are lifted into place by crane
they **come** from every corner
alumni **come** flowing into campus each spring

Table 9. Supplemental training production sentences.

knotty **teak** slowed down the saw
his trusty **teak** saw went missing
gaudy **deak** vests are green
he might veto **deak** findings if necessary

glossy **taupe** shelves went unpainted
her heavy **taupe** suitcase was lost in transit
a pushy **dope** shoved through the line
the sleepy **dope** fell back asleep

an extra **ton** was added to the ship
a rusty **ton** soaked up rainwater
a martini **done** straight up is tasty
the forgery **done** sloppily was obvious

a bikini **top** sold for twenty dollars
a true **top** view was the right angle
suddenly **dop** showed up
radio **dop** surprised the listeners

no **ten** shingles are exactly the same
free **ten** shells from that crack
the new **den** served as a play room
the artsy **den** flooded after the storm

a tepee **too** hastily constructed is not traditional
if you ski **too** frequently it could be dangerous
they **do** have reservations
Fred and Sue **do** share a love for tennis

Table 10. Alveolar generalization production sentences.

elderly **kith** should be respected
bee **kith** hesitate to sting their own
the rabbi **gith** felt forced
cargo **gith** sometimes affects the packages

the baby **coo** seemed very quiet
he will certainly **coo** for attention
pizza **goo** sometimes drips onto the pan
balsa **goo** slid down the tree's trunk

that bayou **curl** stems from a canal
a rodeo **curl** shot is difficult
the new **girl** showed up late
a bubbly **girl** vandalized the school

Sally **can** feel a storm coming
a hobo **can** hitch a ride on a train
go **gan** that roadblock
kiwi **gan** should be discarded

sloppy **car** sometimes causes accidents
a Pakistani **car** swerved to avoid the accident
Thai **gar** fish are not rare
two **gar** females swam by

Barry **came** home from school
a snow flurry **came** from the west
that patio **game** should be easy to play
the pagoda **game** frustrated the kids

Table 11. Velar generalization production sentences.

The Selfish Recruit

They went from the shore through the **town** center, passing a mother with a bratty **tot** screaming loudly. Only minutes before they had hoisted piles of plants over the rusty **keel** section of the boat. Now, their job was to **cull** seaweed from the **kale** spines and leaves. None of them thought one may **taint** the other, so they **took** them from a pile of flora **down** from the docks into a field. In the field, they used pitchforks to sift the **good** from the bad plants. Cleaning every **tine** very well might have made their hard work easier, but they never thought of it. They eventually **got** fed up with the sweaty work, and one by one began to **take** some of their winter clothes off in order to **keep** from overheating. They **could** scarcely **deal** very well with sweating, so they made the **call** for a recruit named Henry to walk two miles to the **town** square to buy all the **talc** from the general store. This fellow had the **gall** first to refuse, saying he did not have a **dime** saved and should not buy **talc** for everyone. In fact, he did not want to make the walk. He threw his grubby **teal** shaded **tuque** fiercely onto a nearby **cot** forgotten in the field a long while ago. It wasn't clear why there was a **cot** spoiling in the wind and sun, but now it had a dirty **tuque** sitting on it.

Someday, Henry may **come** from anger to regret the **time** he ranted and raved. On that day in the field, he was still very bitter. He decided to quickly **take** from each man a **dime** for buying the healing powder. But the **kale** started spoiling in the heat while the men waited until he **took** the **deal** from the leader. Now the men had an additional task: to **cull** seaweed from the **good** sections of the plants to **keep** certain impurities out. These would surely **come** spreading through the plants after they were brought back over the **keel** segment of the boat. The men knew that any impurities would serve to **taint** the whole lot, so they started at the upper fraction and quickly dug very close to the middle. There, they discovered a kind of mold that left a yellow **tint** spreading on each pointy **tine** shining in the sun. At the **time** Henry threw the **teal** hat in anger, some men repairing a nearby **kiln** flue with specialty **caulk** forms began to **gawk** somewhat in disbelief. Previously, the repair men had been talking about weights and measures, saying that a **kip** scale is a standard measure of one-thousand pounds—which is a **tot** heavier than 450 kilograms. They thought the **kiln** should weigh a little more than the **kip** standard. It was late in the day, so they decided to go **down** from their hill to a neighboring city to **dine** hurriedly at a local pub and get dye to **tint** some of the **caulk** for their project. But before they **could** start to **dine** freely at the pub, they **got** sidetracked by the **call** for the recruit’s errand—and then by the **gall** from his initial refusal--so they began to **gawk** fixedly across the field.

6. References

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America* 115: 3171–3183.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance* 10: 340-357.

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

Bedford, F. L. (1995). Constraints on perceptual learning: Objects and dimensions. *Cognition* 54: 253-297

Bloomfield, L. (1927). Literate and illiterate speech. *American Speech* 2(10): 432-439.

Bloomfield, L. (1933). *Language*. New York: Holt.

Bromberger, S., and Halle, M. (2001). The ontology of phonology. in Burton-Roberts, N., Carr, P., and Docherty, G. (eds.). *Phonological knowledge: Conceptual and empirical issues*. Oxford: Oxford University Press, pp. 19-37.

Brown, R. W., & Gilman, A. (1960). The pronouns of power and solidarity. in Sebeok, T. A. (ed.) *Style in language*. Cambridge, MA: MIT Press, pp. 253-276.

Brugmann, K. (1886). *Grundriss der vergleichenden Grammatik der Indogermanischen Sprachen. Kurzgefasste Darstellung der Geschichte des Altindischen, Altiranischen (Avestischen u. Altpersischen) Altarmenischen, Altgriechischen, Lateinischen, Umbrisch-Samnitischen, Altirischen, Gotischen, Althochdeutschen, Litauischen u. Altkirchenslawischen*. Strassburg: Trübner.

Burling, R. (1973). *English in black and white*. London: International Thomson Publishing.

Bybee, J. L. (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Cho, T., & Jun, S-A. (2000). Domain-initial strengthening as featural enhancement: Aerodynamic evidence from Korean. *Chicago Linguistics Society* 36: 31-44.

Cho, T., & Keating, P. (2001). Articulatory and acoustic studies of domain-initial strengthening in Korean. *Journal of Phonetics* 29: 155-190.

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

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT press.

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Lanham, Maryland: University Press of America.

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

Clarke, C. M., & Luce, P. A. (2005). Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. in, McLennan, C. T., Luce, P. A., Mauner, G., & Charles-Luce, J. (eds.). *University of Buffalo Working Papers on Language and Perception*, Vol. 2: 362-366.

Cohn, A., Brugman, J., Crawford, C., & Joseph, A. (2005). Phonetic duration of English homophones: An investigation of lexical frequency effects. *79th Annual Meeting of the Linguistic Society of America*. Oakland, CA.

Cooper, W. E. (1974). Selective adaptation for acoustic cues of voicing in initial stops. *Journal of Phonetics* 2: 303-313.

Cooper, W. E. (1979). *Speech perception and production studies in selective adaptation*. Norwood, New Jersey: Ablex Publishing Corporation.

Diehl, R. L. (1981). Feature detectors for speech: A critical reappraisal. *Psychological Bulletin* 89(1): 1-18.

Dillard, J.L. (1972). *Black English*. New York: Random House.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language* 27: 143–165.

Eimas, P. D., and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4: 99-109.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics* 67(2): 224-238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America* 119(4): 1950-1953.

Fitch, W. T., Hauser, M. D., Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition* 97: 179-210.

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

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language* 49: 396-413.

Fowler, C. A., Levy, E. T., & Brown, J. M. (1997). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language* 37: 24-40.

Gahl, S., and Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80: 748-775.

Gahl, S., and Yu, A. C. L. (2006). Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review* 23: 213-216.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6: 110-125.

Goffman, E. (1955). On face work: An analysis of ritual elements in social interaction. *Psychiatry* 18: 213-231.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition* 22: 1166-1183.

Goldinger, S. D. (1997). Perception and production in an episodic lexicon. in Johnson, K., and Mullennix, J. W. (eds.) *Talker variability in speech processing*. San Diego: Academic Press, pp. 33-66.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105: 251-279.

Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. in *SWAP-2000*: 155-158.

Goldinger, S.D., Pisoni, D. B., & Logan, J.S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17: 152-162.

Guion, S. (1995). Word frequency effects among homonyms. *Texas Linguistic Forum* 35: 103-116.

Hale, M., & Reiss, C. (2001). Phonology as cognition. in Burton-Roberts, N., Carr, P., and Docherty, G. (eds.). *Phonological knowledge: conceptual and empirical issues*. Oxford: Oxford University Press, pp. 161-184.

Hauser, M. D., Chomsky, N., Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve?. *Science* 298:1559-1579.

Harris, L. B. (1980). *An assessment of current hypotheses concerning selective adaptation using phonetic stimuli*. unpublished MA thesis, State University of New York at Binghamton.

Harris, Z. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.

Harris, Z. (1962). *String analysis of sentence structure*. The Hague: Mouton.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics* 31: 373-405.

Hintzmann, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review* 93: 411-428.

Hock, H. H. (1986). *Principles of historical linguistics*. The Hague: Mouton de Gruyter.

Hooper, J. B. (1976). *An introduction to natural generative phonology*. New York: Academic Press.

Howes, D. H. (1954). On the interpretation of word frequency as a variable affecting speech recognition. *Journal of Experimental Psychology* 48: 106-112.

Howes, D. H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America* 29: 296-305.

Idsardi, W. J. (2006). A simple proof that Optimality Theory is computationally intractable. *Linguistic Inquiry* 37: 271-275.

Ingemann, F. (1977). *Speech synthesis by rule using the FOVE program*. Paper presented at the International Congress of Phonetic Sciences. Miami, FL.

Jakobson, R., Fant, G., & Halle, M. (1952/1963). Preliminaries to speech analysis: The distinctive features and their correlates. Cambridge, MA: MIT Press.

Jesperson, O. (1924). *The philosophy of grammar*. London: Allen and Unwin.

Johnson, K. (1990). Contrast and normalization in vowel perception. *Journal of Phonetics* 18: 229-254.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. in Johnson, K., and Mullennix, J. W. (eds.) *Talker variability in speech processing*. San Diego: Academic Press, pp. 145-165.

Johnson, K. (2005). Speaker normalization in speech perception. in Pisoni, D.B. & Remez, R. (eds.). *The handbook of speech perception.* Oxford: Blackwell Publishers, pp. 363-389.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. in Sole, M.J., Beddor, P. & Ohala, M. (eds.). *Experimental approaches to phonology. In honor of John Ohala*. Oxford: Oxford University Press, pp. 25-40.

de Jong, K.J. (1995). On the status of redundant features: The case of backness and roundness in American English. in Connell, B., and Arvaniti, A. (eds.). *Papers in laboratory phonology* 4. Cambridge: Cambridge University Press, pp. 68-86.

de Jong, K. J. (2001a). Rate-induced resyllabification revisited. *Language and Speech* 24: 197-216.

de Jong, K.J. (2001b). Effects of syllable affiliation and consonant voicing on temporal adjustment in a repetitive speech production task. *Journal of Speech, Language, and Hearing Research* 44: 826 - 840.

de Jong, K.J. (2003). Temporal constraints and characterizing syllable structuring. in Local, J., Ogden, R., & Temple, R. (eds.). *Papers in laboratory phonology 6*. Cambridge: Cambridge University Press, pp 253 - 268.

de Jong, K.J., Lim, B. J., & Nagao, K. (2004). The perception of syllable affiliation of singleton stops in repetitive speech. *Language and Speech* 47: 241 - 266.

Kat, D., & Samuel, A. G. (1984). More adaptation of speech by nonspeech. *Journal of Experimental Psychology: Human Perception and Performance* 10(4): 512-525.

Kraljic, T., & Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology* 51: 141-178.

Kraljic, T., & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review* 13(2): 262-268.

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

Labov, W. (1963). The social motivation of a sound change. *Word* 19: 273-309.

Labov, W. (1966). *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, W. (1969). Contraction, deletion and inherent variability of the English copula. *Language* 45: 715-762.

Labov, W. (1972) . *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1981). Resolving the Neogrammarian controversy. *Language* 57: 267-308.

Lakoff, G. (1973). Fuzzy grammar and the performance/competence terminology game. *Chicago Linguistics Society* 9: 271-291.

Lakoff, R. T. (1973). Language and women's place. *Language in Society* 2: 45-80.

Lancelot, C., & Arnauld, A. (1660). *Grammaire général et raisonnée.* Paris: Pierre le Petit. Translated into English as *A general and rational grammar containing the fundamental principles of the art of speaking, explained in a clear and natural manner. With the reasons of the general agreement, and the particular differences of languages translated from the French of Messieurs de Port Royal* (1753). London: J. Nourse.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology* 140.

Lisker, L., & Abramson, A. S. (1971). Distinctive features and laryngeal control. *Language* 47(4): 767-785.

Maye, J., Aslin, R., & Tannenhaus, M. (2003). In search of the weckud wetch: Online adaptation to speaker accent. in *Proceedings of the 16th Annual CUNY Conference on Human Sentence Processing*, March 27-29, Cambridge, MA.

McCarthy, J. J. (2002). *A thematic guide to Optimality Theory*. Cambridge, UK: Cambridge University Press.

Mencken, H. L. (1919). *The American language*. New York: Alfred A. Knopf.

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American Sociological Review* 1: 894-904.

Merton, R. K. (1949). *Social theory and social structure*. New York: The Free Press.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica* 43: 106-115.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics* 46: 505-512.

Munson, B. (in press). Lexical access, lexical representation, and vowel production. in, Cole, J. S. & Hualde, J. I. (eds.). *Papers in laboratory phonology* 9. New York: Mouton de Gruyter.

Newmeyer, F. J. (2003). Grammar is grammar and usage is usage. *Language* 79: 682-707.

Newmeyer, F. J. (2006). On Gahl and Garnsey on grammar and usage. *Language* 82(2): 399-404.

Nielsen, K. Y. (2005). Generalization of phonetic imitation across place of articulation. *Proceedings for ISCA Workshop on Plasticity in Speech Perception*. University College London, London, UK.

Nielsen, K. Y. (2006). Specificity and generalizability of spontaneous phonetic imitation. *Proceedings of the 9th International Conference on Spoken Language Processing*. Pittsburgh, USA.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology* 47: 204-238.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115: 39-57.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report 10*, Indiana University, 357-376.

Ohde, R. N., & Sharf, D. J. (1979). Relationship between adaptation and the percept and transformations of stop consonant voicing: effects of the number of repetitions and intensity of adaptors. *Journal of the Acoustical Society of America* 66(1): 30-45.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119: 2382-2393.

Phillips, B, S. (1984). Word frequency and the actuation of sound change. *Language* 60: 320-342.

Picheny, M.A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research* 28: 96-103.

Pierce, C. S. (1867). On a new list of categories. *Proceedings of the American Academy of Arts and Sciences* 7: 287-298.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. in Bybee, J. and Hopper, P. (eds.). *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins: 137-157.

Pierrehumbert, J. (2002). Word-specific phonetics . *Laboratory phonology VII*, Berlin: Mouton de Gruyter, 101-139.

Pierrehumbert, J. (2003). Probabilistic phonology: discrimination and robustness. in Bod, R., Hay, J., and Jannedy, S. (eds.). *Probability theory in linguistics*. Cambridge, MA: The MIT Press, pp. 177-228.

Pisoni, D. B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. *Proceedings of the 1990 International Conference on Spoken Language Processing*. Kobe, Japan, pp. 1399-1407.

Pisoni, D. B. (1997). Some thoughts on normalization in speech perception. in Johnson, K., and Mullennix, J. W. (eds.) *Talker variability in speech processing*. San Diego: Academic Press, pp. 9-32.

Pisoni, D. B. & Tash, J. (1975). Auditory property detectors and processing place features in stop consonants. *Perception & Psychophysics* 18(6): 401-408.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception and Performance* 19: 1-27.

Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language* 81(4): 927-964.

Porter, R. J., Jr., & Castellanos, F. X. (1980). Speech-production measures of speech perception: rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America* 67(4): 1349-1356.

Porter, R. J., Jr., & Lubker, J. F. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research* 23: 593-602.

Remez, R., Cutting, J., & Studdert-Kennedy, M. (1980). Cross series adaptation using song and string. *Perception & Psychophysics* 27: 524-530.

Rosen, S. M. (1979). Range and frequency effects in consonant categorization. *Journal of Phonetics* 7: 393-402.

Ross, J. R. (1973). Slifting. in *The formal analysis of natural languages: Proceedings of the First International Conference*. Gross, M., Halle, M., and Schuetzenberger, M. P. New York: Mouton, pp. 133-169.

Sampson, G. (1974). Is there a universal phonetic alphabet? *Language* 50(2): 236-259.

Samuel, A. G. (1986). The role of the lexicon in speech perception. in, Schwab, E. C., & Nusbaum, H. C. (eds.). *Perception of speech and visual form: theoretical issues, models, and research*. New York: Academic Press.

Samuel, A. G., & Newport, E. L. (1979). Adaptation of speech by nonspeech: evidence for complex acoustic cue detectors. *Journal of Experimental Psychology: Human Perception and Performance* 5(3): 563-578.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics* 25: 421-436.

Sapir, E. (1921). *Language: an introduction to the study of speech*. New York: Harcourt, Brace, and World.

Sato, M., Schwartz, J-L., Abry, C., Cathiard, M-A., & Lœvenbruck, H. (2006). Multistable syllables as enacted percepts: a source of an asymmetric bias in the verbal transformation effect. *Perception & Psychophysics* 68(3): 458-474.

de Saussure, F. (1916). *Cours de linguistic général*. Paris: Payot.

de Saussure, F. (1996). The object of linguistics. in Cobley, P. (ed.) *The communication theory reader*. London: Routledge, pp. 37-47.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America* 35: 200-206.

Schuchardt, H. (1885). Über die lautgesetze: gegen die junggrammatiker. Berlin: Verlag von Robert Oppenheim, pp. 1-39.

Scobbie, J. M. (2005). The phonetics phonology overlap. in Scobbie, J. M., Mennen, I., Watson, J. (eds.). *QMUC speech science research centre working papers WP1*.

Shockley, K., Sabadini, L. & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics* 66: 422-429.

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

Smiljanic, R., & Bradlow, A. R. (2005b). Does clear speech enhance the voice onset time contrast in Croatian and English? *Journal of the Acoustical Society of America* 118(3): 1900.

Stetson, R. H. (1951). *Motor phonetics: A study of speech movements in action*. Amsterdam: North-Holland.

Stevens, K. N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. in Denes, P. B., David, E. E., Jr. (eds.). *Human communication: a unified view*. New York: McGraw Hill, pp 51-66.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics* 17: 3-46.

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.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance* 7: 1074-1095.

Tartter, V. C., & Eimas, P. D. (1975). The role of auditory feature detectors in the perception of speech. *Perception & Psychophysics* 18(4): 293-298.

Tennpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review* 2: 339-363.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustic Society of America* 92: 723-735.

Wang, W. S-Y. (1969). Competing changes as a cause of residue. *Language* 45: 9-25.

Warren, R. M. (1961). Illusory changes of the distinct speech upon repetition—the verbal transformation effect. *British Journal of Psychology* 52(3): 249-258.

Warren, R. M. & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology* 71(3): 612-613.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science* 167: 392-393.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The Linguistic Review* 23: 247-274.

Wolfram, W. (1969). Linguistic correlates of social differences in the negro community. in Alatis, J. (ed.) *Georgetown monograph series on languages and linguistics* No. 22, pp. 249-257.

Curriculum Vitae

Mark VanDam was born in Lansing, Michigan, on September 6, 1977. He received a Bachelor of Arts in German and English from Calvin College in 1999, and a Master of Arts in Linguistics from Indiana University in 2003. Following completion of his doctoral work, he will begin a post-doctoral appointment at Boys Town National Research Hospital in Omaha, Nebraska.