UNIVERSITY OF CALIFORNIA

SANTA CRUZ

THE FILTERING LISTENER: DISPERSION IN EXEMPLAR THEORY

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF ARTS

in

LINGUISTICS

By

Thomas Denby

June 2013

The Thesis of Thomas Denby
is approved:

Professor Jaye Padgett, co-chair

Assistant professor Grant McGuire, co-chair

Assistant professor Pranav Anand

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

ABSTRACT

THE FILTERING LISTENER: DISPERSION IN EXEMPLAR THEORY

THOMAS DENBY

Phonetic dispersion has been proposed as an explanation for a number of sound-change phenomena, including vowel chain shifts, compensatory sound change, and universal trends within phoneme inventories. These explanations usually take the form of speaker-based accounts (principally H&H theory; see Lindblom, 1986). But there is a lack of empirical evidence for speaker-based approaches (McGuire and Padgett, 2011), which have also been criticized as being teleological (e.g. Blevins, 2004). This thesis explores an alternative, listener-based, perceptual account of dispersion: the Filtering Listener (Labov, 1994: 587; Wedel, 2006; McGuire and Padgett, 2011). Based in exemplar theory (Nosofsky, 1986; Pierrehumbert, 2001; etc.), the Filtering Listener hypothesis argues that when listeners are confronted with phonetically ambiguous percepts, they may not store them to phonetic memory. In turn, these unstored percepts do not update the phonemic categories of the listener, and are thus not reflected in that listener's future productions. This thesis explores the Filtering Listener as a mechanism for contrast maintenance and source of sound change, and tests its predictions experimentally in a perceptual

recognition task. Participants heard phonetically ambiguous target words and phonetically unambiguous control words in noise and were asked to identify them, following Goldinger (1996). This task was repeated in 4 identical experimental blocks. Results showed that participants improved their accuracy of recognition much more over the course of the experiment for the unambiguous condition, suggesting that their storage of ambiguous percepts was degraded in memory. This provides promising preliminary evidence for the Filtering Listener; further replication, with more tightly controlled pilot studies, is required.

ACKOWLEDGMENTS

First and foremost, I'd like to thank my advisors, Grant McGuire and Jaye Padgett, whose limitless support has made this project possible. Jaye has acted as a guide and a friend from from the moment I arrived at UC Santa Cruz; I have no doubt his support over the last two years has been instrumental in my development as a linguist. Thanks to Grant for sparking my love of phonetics and teaching me most of what I know about speech sounds and experimental design. It has truly been a pleasure to work with both of them. Thanks also to my final committee member, Pranav Anand, whose constructive and challenging questions have done so much to shape this project for the better.

I would also like to thank the members of the 2013 Winter Research Seminar at UCSC, as well as the members of Jaye Padgett's 2012 Fall Phonology Seminar, whose insights proved invaluable in the early stages of this project. Thanks also to Alyssa Heekin, Karl DeVries, and Adrian Brasovaneau for all their help with statistics, experimental design, and presentation. Thanks as well to Matt Wagers, for his extremely useful advice on data analysis, and, in particular, Adam Morgan, who was an

illimitable source of late-night moral support and sage advice on experimental design.

I would also like to thank Ben Munson, Andries Cotzee, Janet Pierrehumbert, and, in particular, Andy Wedel for their comments and encouragement. They played a crucial part in the formation of this project, as well as convincing me that it was worthwhile in the first place.

Finally, I would like to thank my family, without whose support I wouldn't be here (or anywhere, for that matter). In particular, thanks to my mom for cheerfully humoring me countless times as I thought aloud about dispersion, exemplar theory, and experimental design.

1. Introduction

Linguistic dispersion is principally defined as the maintenance of sufficient pyscho-acoustic distance between phonemic categories within a given perceptual space. Its effects have been claimed to surface in a number of closely related phenomena, including maintenance of phonemic contrast; universal trends within phoneme inventories; compensatory sound changes and phonologization; and phonetic shift of phonemic categories in response to sound change, as in vowel chain shifts.

From a functionalist perspective, dispersion is no mystery: contrast is at the heart of phonology, and without the separation of phonemic categories language wouldn't be possible. Informally, one might imagine phonemes "wanting" to maintain the distance between one another. In so doing, they would be maintaining perceptual distinctiveness of phonemes and, ultimately, the successful exchange of information between speakers. Of course, phonemes are not rational agents, and this is merely description, not explanation.

The challenge, then, is to propose a mechanism by which phonemic categories maintain phonetic contrast. In this paper, I propose a listener-

based mechanism: the Filtering Listener (Labov, 1994: 587; Wedel, 2006; see McGuire and Padgett, 2011 for terminology and discussion). The Filtering Listener (FL) hypothesis argues that when listeners are confronted with phonetically ambiguous percepts, they may not store them to phonetic memory. In turn, these unstored percepts do not update the phonemic categories of the listener, and are thus not reflected in that listener's future productions. Over time, this filter reduces the number of phonetic traces stored between phonemic categories relative to those close to the center of the distribution, providing a buffer between categories, resulting in dispersion. Wedel (2006) has demonstrated that this filter can be implemented in a self-organizing model of the perception-production loop to simulate dispersion.

In the second section of this paper, I couch the FL hypothesis within the framework of exemplar theory (Johnson, 1997; 2007; Garrett and Johnson, 2011; Pierrehumbert, 2001; 2002; Wedel, 2004; 2006), identifying two possible routes to dispersion. In the third section, I briefly compare speaker- and listener-based accounts for dispersion, and survey the evidence for the FL. In the fourth, I describe the evidence for dispersion within the phenomena mentioned above, and explore how the FL does or does not predict the observed patterns. In the fifth section, I share the

results of an experiment run for the present paper in which participants identified phonetically ambiguous target stimuli and phonetically unambiguous controls in noise. The accuracy of stimulus identification was tabulated over the four phases of the experiment, with improvements in accuracy presumably reflecting successful exemplar storage (following Goldinger, 1996). The results from this experiment suggest that exemplar storage is degraded for phonetically ambiguous stimuli, and that a mechanism such as the Filtering Listener could account for these results. Finally, I discuss the results and suggest future directions for exploring the FL hypothesis in section 6.

2. Exemplar Theory

Exemplar models of speech production and perception (Garrett and Johnson, 2011; Johnson, 1997; 2006; 2007; Pierrehumbert, 2001; 2002; Wedel, 2004; 2006; etc.) have gained considerable traction over the last decade. Sometimes referred to as episodic or instance-based models, exemplar theory is essentially a usage-based model of memory and categorical learning originally adapted from psychology (see Bruner, Goodnow, & Austin, 1967 and Nosofsky, 1986) in which percepts are stored as individual exemplars. For speech perception, this means each time a word is heard it is stored as its own detailed exemplar, possibly containing linguistic and socio-linguistic information about the speaker. "Clouds" of these exemplars then collectively make up individual categories—that is, either words or phonemes. The correct level of representation for exemplars is not immediately apparent. I will assume, following Pierrehumbert (2002), that exemplars are stored both at the lexical and phonemic level, although for the present study it is crucial only that they be stored at the lexical level.

The dispersion effect falls out from exemplar theory by two means, both involving ambiguous utterances. An ambiguous token, by definition, has a

higher than normal chance to be miscategorized, in which case it will not affect the exemplar cloud of the intended word. In this way, ambiguous tokens will be stored at a lower rate for any given category, and over the course of thousands of simulated iterations of the production/perception loop, categories separate (see Wedel, 2004b; 2006 for discussion). In the strong dispersion hypothesis, Wedel (2006) adds a stipulation to the model that builds on this: every time an exemplar is categorized, there is a small possibility that ambiguous tokens are sometimes not categorized at all, and thus not stored as exemplars. In other words, that there is something intrinsically difficult about ambiguous tokens to store in memory. The two hypotheses will be explored in further detail in this section.

In this section I discuss the origins of exemplar theory within speech perception; its predictions with regards to frequency and density effects; the mechanism by which categories are formed; and the necessity for the Filtering Listener due to cross-category blending inheritance.

2.1 An episodic memory

Variability in the speech signal is a well-documented phenomenon. The production of any word is dependent on the unique vocal properties of its speaker, as well as the context within which it is uttered. Thus speaking

rate, prosody, and phonetic context all shape the phonetic realization of any given word. But despite this variability, listeners are remarkably successful in decoding the speech signal (see section 4.2 for a more complete discussion).

This fact is sometimes cited as evidence for words being represented canonically in memory, broken up into discrete abstract categories, as in modular feed-forward models. In one version of this paradigm, when a word is heard it is sent to a phonological buffer, where it is normalized across speakers and phonetic contexts, and matched to a syntactic/semantic lexical entry. Surface phonetic details—i.e. those that account for variability—are filtered out by the buffer and not stored in memory. This allows for efficiency in conversation, as the semantic and syntactic attributes of words are easily retrieved while phonetic detail is used for perception, but not ultimately stored in long-term memory. (See Pierrehumbert, 2002, for discussion.)

Goldinger (1996, 1998, 2000), however, found that speaker variability affects memory in a number of different tasks. In one such task (Goldinger, 1996), participants were exposed to words in a learning phase, and then

given the same words with an equal number of new words in the test phase. The task in the test phase was to say whether the word was "old" or "new" (i.e. heard in the learning phase or not). Participants were more accurate if the stimulus was spoken by the same speaker in both the learning and testing phase. Goldinger (1996) considers this evidence that phonetic detail is stored in memory. Furthermore, in a shadowing-task, Goldinger (1998) found that closeness of imitation is negatively correlated with frequency. This suggests that fine-grained phonetic detail must be stored in memory, since it is utilized—in the form of knowledge of the speaker's voice—in memory tasks. This led Goldinger to argue that for an episodic (or exemplar) theory of memory, in which individual "traces" are stored with phonetic detail, and "collectively represent individual words" (Goldinger, 1996: 1166).

2.2 Frequency and density effects

Crucial to any exemplar-based account are frequency and neighborhood density. If a high-frequency word is heard, a large number of exemplars will be activated. Even if the surface phonetic details of the token perfectly match a significant number of exemplars—if it is spoken in a familiar voice, for example—the "bonus" to recall from this matching will be drowned out by a generic activation of many exemplars. If, however, the word is low-frequency, voice and other effects should be significant,

since the number of exemplars matching in surface detail should make up a larger ratio of the total exemplar cloud. Goldinger (1998) and Goldinger (2000) demonstrated this convincingly, in imitation tasks followed by AXB discrimination tasks. In Goldinger (2000), participants recorded baseline tokens of words, which were manipulated for frequency. The next day, speakers heard these same words in a study period, where the number of exposures was manipulated. Five days later, participants recorded the same words as on the first day—these were the test tokens. In a follow-up experiment, a different set of participants were given an AXB discrimination task, in which A and B were the baseline and test tokens, and X was the original recording heard during the study period. In other words, did participants from the first study change their productions after the study period, and imitate what they heard? The results strongly indicated that this was the case, with both frequency and number of exposures correlated with closeness of imitation.

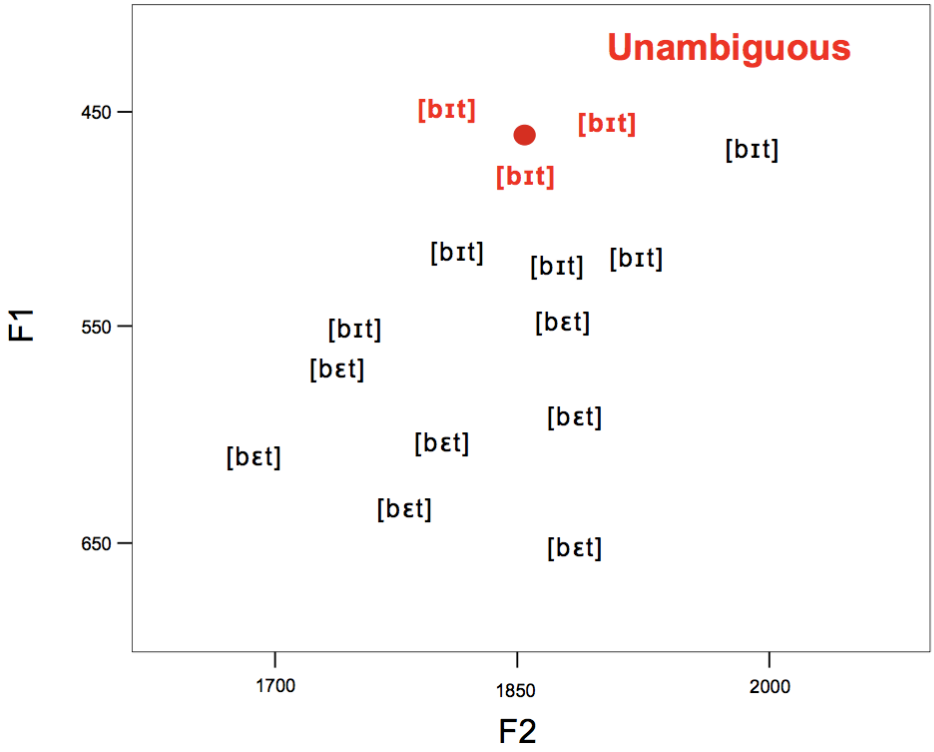
Munson and Solomon (2004) attempted to disentangle frequency and neighborhood density in a production study, which were confounded in Goldinger's studies, as well as earlier studies on "hard" and "easy" words (Wright, 1997; 2004). Munson and Solomon (2004) found that hyper-articulation (i.e. expansion of the vowel space) occurred independently for

both low-frequency words and high-density words, and that words which were both low-frequency and high-density ("hard" words) showed the greatest effect.

2.3 Ambiguous tokens and dispersion

In exemplar models of perception, whenever a word is perceived, its auditory properties are compared with those of other exemplars in the perceptual space through a similarity function. The more similar a stored exemplar is with the incoming token, the higher its activation level. The sum of activation of all exemplars for any given category determines how likely it is that the incoming token will be categorized as an instance of that category. In other words, the category with the highest sum activation "wins". The incoming stimulus is categorized and stored, and the exemplar space is modified, which is reflected in future productions. In Figure 1, we can see, in an idealized exemplar space, a lexical exemplar being stored with an unambiguous vowel. Note that the stimulus activates only one.

Figure 1: Idealization of unambiguous exemplar storage; exemplar represented by red dot; activation represented by red highlighting of previously stored exemplars.



This mechanism allows for gradient shifting of phonetic categories based on exposure, something that is difficult to achieve with a model that normalizes incoming speech and matches it to discrete, abstract lexical entries. Furthermore, in a simulation of the production/perception loop, each time a production is chosen its output acts as the next incoming stimulus. This allows for an exemplar space originally "seeded" with only one exemplar to dynamically self-organize as the production/perception

loop is carried out thousands of times, with an exemplar added with each iteration.

The first hurdle for exemplar models is *intra-category* entrenchment. A crucial part of Pierrehumbert's (2001; 2002) model is the addition of some random noise with the production of each exemplar. This is meant to reflect natural variation in production, and since the model starts with 1 exemplar, noise is required if any difference between exemplars is to develop at all. If there is noise added with the production of every new exemplar, however, the distribution for any given category will eventually flatten out, approaching a Gaussian distribution with enough iterations of the model. Pierrehumbert (2001; 2002) also adds a lenition (or hypo-articulation) bias, hoping to capture work by Bybee (2004) showing that higher frequency words are more like to show lenition. If a lenition bias is also included with the production of each exemplar, its additive effects are similar to those of random variation: it only increases the flattening of the distribution over time.[1]

[1] This works nicely, however, to capture Bybee's (2004) frequency effects. Since lenition is additive, the more times you use a word the more lenition will be able to operate on it. This is only important in this discussion because it provides some evidence that exemplar representations should exist at the lexical level (not to the exclusion of the phonemic level, however.)

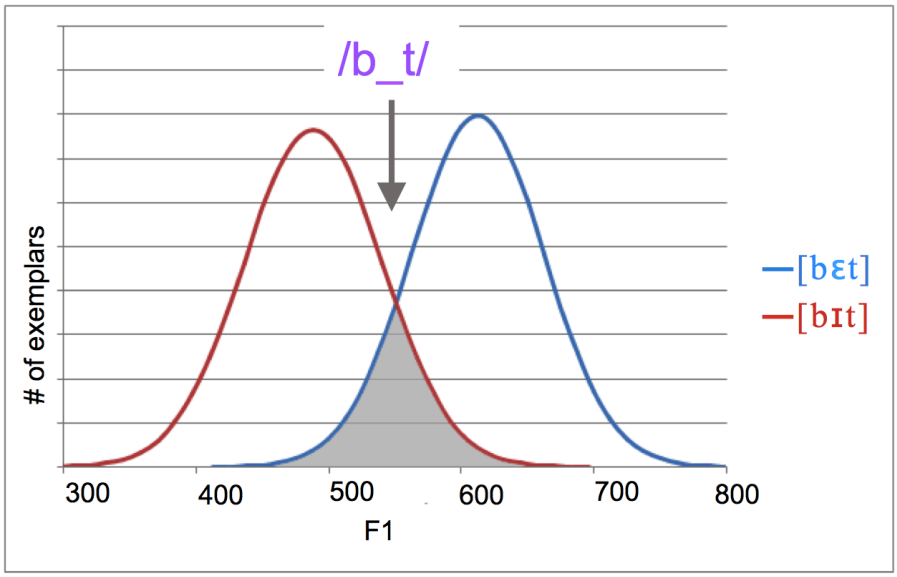
Pierrehumbert's resolution of this problem is to change the way exemplars are chosen for production. Instead of randomly choosing a single exemplar from any given category for production, an entire neighborhood (using a window function) is averaged around that single exemplar. That average is then sent to production. This type of "blending inheritance" (Wedel, 2006) causes a reversion to the mean, which attenuates the flattening effect of noise and lenition. Entrenchment is, of course, necessary for the preservation of contrast between phonemic categories.

2.4 Weak and strong hypotheses

In the model described above, ambiguous tokens are at a selective disadvantage. While the chance that an unambiguous token is categorized as intended by the speaker is close to ceiling, an ambiguous token, by definition, has a diminished chance of being categorized as intended. A perfectly ambiguous token, i.e. one whose phonetic properties have it perfectly spaced between two mean exemplar distributions, has a 50% chance of being attributed to either category. Figure 2 depicts a hypothetical listeners' distribution of two neighboring categories, [bɛt] and

[bɪt]. Note that the distributions are overlapping, and the incoming stimulus has an equal chance of being stored as either word.

Figure 2: Idealized storage of perfectly ambiguous exemplar on the F1 scale between [bɛt] and [bɪt]



To illustrate this further, let's say we have two hypothetical categories, A and B, with overlapping distributions, and we add an equal number of tokens—let's say 20—at three points in the exemplar space: the mean distribution of A, the mean distribution of B, and a perfectly ambiguous area in between the two. All 20 unambiguous tokens added to mean distribution of A will be categorized as A and 20 as B, and the 20 ambiguous tokens will tend to split 50/50. If we now just look at category A, it will have 20 unambiguous tokens added in the mean distribution, and only 10 added from the ambiguous area. The next production target,

then, has a greater chance of being selected from the unambiguous part of the exemplar space. Over the course of many exposures this disadvantage is exaggerated, resulting in the mean distribution of both categories moving away from one another and an increasingly smaller chance that an ambiguous exemplar will be chosen as a production target. This is the weak hypothesis of listener-based dispersion in exemplar theory. Labov (1994:586-587) sums up the assumption underlying this process: "Misunderstood tokens may never form part of the pool of tokens that are used", so that if a listener "fail[s] to comprehend [a] word and the sentence it contains...this token will not contribute to the mean value" of the target segment. If the weak hypothesis can drive dispersion with exemplar models, why is the Filtering Listener necessary at all?

As Wedel (2004b; 2006) notes, there may also be effects of *cross-category* blending inheritance—that is, a reversion to the mean across the entire exemplar space. Wedel (2006) suggests this may occur by two means. First, certain dynamical attractor models of production (Kelso et al., 1992; Kelso, 1995, as cited in Wedel, 2006) have attractors that "systematically bias motor output toward previously practiced outputs in relation to similarity (Zanone and Kelso, 1992; 1997)" (Wedel, 2006: 18). This bias towards previous outputs creates a reversion to the mean, and if it is allowed to

operate over different phonemic categories, "nearby attractors will show a tendency to merge over time, resulting in increasing overlap of the contents of neighboring phonological categories" (Wedel, 2006: 18).

The other possible cause of cross-category blending inheritence comes from the perceptual side: the perceptual magnet effect. As the acoustic features of a stimulus approach the "best exemplar" (Iverson and Kuhl, 2000) of a given phoneme, differences become less noticeable. This is a reflection of the distortion of the underlying perceptual space around phonemes. In some models of perception (Guenther and Gjaja, 1997), percepts are subject to this warping before any categorization has taken place. This leads to a system-wide reversion to the mean along any given phonetic dimension, since percepts are warped towards the center of the distribution along that dimension, regardless of category boundaries. A third source of cross-category reversion to the mean, at least among vowels, might come in the form of a consistent production bias towards centralization, as assumed by speaker-based models (e.g. Lindblom, 1986).[2] One could imagine the effects of this being additive, applying each time a vowel is produced.

[2] Thanks to Jaye Padgett for suggesting this.

The possibility of cross-category blending suggests that a stronger hypothesis may be needed (although it also does not rule out the weak hypothesis). To solve this problem, Wedel (2006) suggests a consistent bias, either in perception or production, that would, in effect, not treat all exemplars as equal. In other words, there needs to be a mechanical selective pressure against ambiguous exemplars. In production, this might come in the form of hyper-articulation (e.g. Lindblom, 1986; see Section 3.1 for further discussion), which biases speakers to produce exemplars at the extremes of the distribution (as long as no other category is close to that extreme).[3] Another selective bias would be the Filtering Listener, which Wedel (2006) utilizes in his model.

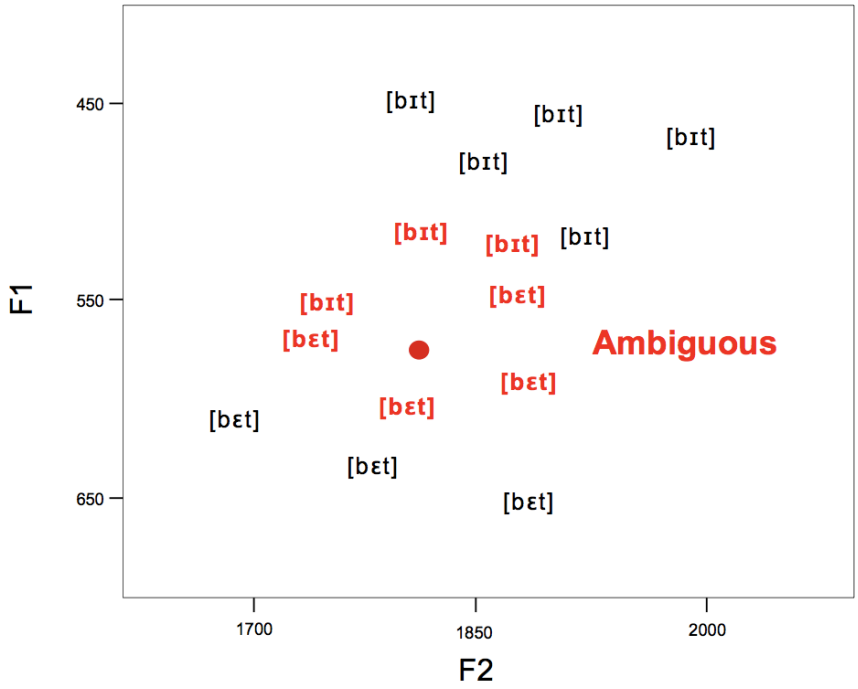
In Pierrehumbert's (2001) model, there is a very weak form of this selective bias. When competition between categories reaches a certain point, exemplars are not stored at all: "If an incoming stimulus is so ambiguous that it can't be labeled, then it is ignored rather than stored. That is, the exemplar cloud is only updated when the communication was successful to the extent that the speech signal was analyzable" Pierrehumbert (2001: 13). In other words, "In the case of a tie [between

[3] For more on the distinction between speaker- and listener-based accounts of dispersion, see section 3.

two categories during categorization], the utterance is discarded" (Pierrehumbert, 2001: 16). The number of exact ties, however, likely is not a strong enough bias to retard cross-category blending.

Wedel (2006, p. 25) expands the possibility of an ambiguous exemplar being dropped, adding a stipulation to "allow occasional, stochastic failure of categorization, where the probability of failure is greater to the degree that exemplar activation is distributed among multiple lexical categories." In other words, there is something inherent to ambiguous tokens that makes them difficult to store in memory. This is the strong dispersion hypothesis—that is, the FL. We can see an idealized version of a stimulus that activates multiple categories very strongly in Figure 3.

Figure 3: Idealized storage of ambiguous exemplar; the red dot represents the stimulus



But even this strong hypothesis only allows a small chance that ambiguous tokens won't be stored as exemplars. Wedel (2006) sets the upper limit of this chance—i.e. for a perfectly ambiguous token—at only 10%. Even so, that number is high enough that we would expect to be able to find experimental evidence for the strong hypothesis.

The focus of this study was to test the strong dispersion hypothesis: Does an ambiguous token have a chance of not being stored at all in the

exemplar space? In this vein the method of Goldinger (1996) was followed, while changing the key manipulation from voice to ambiguity.

That said, this experiment cannot provide evidence for the weak hypothesis, which relies on only two very basic assumptions: first, that ambiguous tokens are sometimes miscategorized (which is, by definition, true); and second, that exemplars from one category cannot affect the distribution of another category, even if those categories have overlapping distributions (per Labov, 1994). If we assume these two things, as well as the activation-weighted averaging over a group of exemplars that Pierrehumbert (2001) proposes to derive entrenchment, then there is no experimental question to propose—dispersion is achieved gradually through a statistically driven process, and ambiguous tokens are not intrinsically degraded in memory.

3. Speaker- vs. Listener-based accounts

3.1 The considerate speaker

While the FL models dispersion as a perceptual mechanism, most previous accounts been centered on production rather than perception. The most notable example is H&H theory (see Liljencrants and Lindblom, 1972; Lindblom, 1986; Lindblom, 1990), which attempts to explain dispersion as a result of hyper- and hypo-articulation. This is an explicitly teleological explanation, with the speaker having "tacit awareness of the listener's access to sources of information independent of the signal and his judgment of the short-term demands for explicit signal information" (Lindblom, 1986: 403). One of the often-cited (Blevins, 2004; Ettlinger, 2007; Wedel, 2006; etc.) benefits of exemplar theoretic, listener-based accounts is that they do not have to assume speaker or listener intention, tacit or not—instead, dispersion effects fall out from the categorization mechanisms described in the previous section. Putting aside these important but ultimately philosophical issues, which of the two accounts can we make a stronger empirical case for?

One of the pillars of H&H theory is the observation that speakers "tune their performance according to communicative and situational demands"

(Lindblom, 1986: 403). This is undeniably true, and has many manifestations: clear speech (Uchanski, 2005); child-directed speech (Fernald and Kuhl, 1987; Burnham et al., 2002); foreign-directed speech (Uther et al., 2007; Van Engen et al., 2010); speech countering adverse listening conditions (Hazan and Baker, 2011); and even pet-directed speech (Burnham et al., 2010). In particular, the fact that clear speech commonly leads to a hyper-articulated vowel space (e.g. Moon and Lindblom, 1994; Bradlow et al., 2003) seems to suggest that speaker-based effects may be driving dispersion. Putting aside the issue that clear speech may in part be an artifice of laboratory settings, this does not in itself provide evidence for a speaker-based account. There is no causal link between the types of across-the-board, global effects that arise in clear speech and many of the dispersion effects to be described in the following section, such as chain shifts and compensatory sound change. These types of sound change seem to be triggered when specific contrasts are in danger of being erased, and are unlikely to be caused by global factors.

Moreover, this does not necessarily distinguish between online, speaker-based accounts and offline, listener-based accounts. Hyper-articulation in clear speech might be explained in listener-based accounts as a prototype effect (see Boersma and Hamann, 2008; Blevins, 2004: 285-289;

Pierrehumbert, 2001). This refers in part to the finding that tokens with extreme, and thus unlikely, productions can still be judged as better examples of a particular phoneme than their less extreme counterparts (Johnson, Flemming, and Wright, 1993). It has been suggested (Blevins, 2004) that when speakers intend to produce clear speech they may fall back on these prototypical exemplars.

More relevant evidence for speaker-based accounts than global clear speech effects would be speakers making *local* adjustments in production to avoid ambiguities. For example, hyper-articulating a particular phoneme that may be causing lexical ambiguity. Such an effect would provide causal link to dispersion effects equally as strong as that of the FL.[4] McGuire and Padgett (2011) tested this local, speaker-based dispersion effect, referring to it as the Considerate Speaker (CS). Both the CS and prototype theory predict that a corner vowel will be disambiguated through hyper-articulation. In the case of non-corner vowels, however, these theories make differing predictions. Let's take the example of [pɪn], [pɛn], and [pæn]. If a speaker is in a context in which she has to disambiguate [pɛn] for the listener, does it matter what she is trying to

[4] While I am portraying these as competing theories, the FL and CS are not mutually exclusive—both can be modeled in exemplar theory, and it is possible that both are contributing to dispersion effects.

disambiguate it from? If she is considerate of her speaker, we might expect her to produce [ɛ] with a higher F1 when trying to disambiguate it from [ɪ], and a lower F1 when disambiguating it from [æ]. If she relies on her prototype, on the other hand, she will simply produce a uniquely [ɛ]-like vowel, regardless of context.

McGuire and Padgett (2011) tested this in a production study, in which participants were told they were assisting the experimenters in testing a speech-recognition program. Participants were elicited for words like *pen*, which had minimal pairs for [ɪ] and [æ]. The computer would then try to "understand" participants' productions; the participant would correct the computer with a second production if it reported "hearing" the wrong word. The CS predicts that in cases in which the computer incorrectly reported hearing [æ], the participant would raise the vowel in their repetition of the word to disambiguate it, and vice versa with [ɪ]. The experimenters found some global clear speech effects (longer vowel durations), but no evidence of local dispersion whatsoever. These results are consistent with the prototype effect.

3.2 Evidence for the Filtering Listener

Beyond its usefulness in modeling contexts, there is some experimental evidence for the Filtering Listener.[5] Wedel (2006) cites the predictions of the Neighborhood Activation Model (NAM) of Luce and Pisoni (1998). NAM makes similar predictions to exemplar theory in terms of lexical activation—stimuli that activate only one lexical category will be accurately and quickly perceived. Stimuli that activate multiple lexical categories, however, will result in difficulty in processing time—e.g. reduced accuracy and longer reaction times in a lexical decision task. If the competition is strong enough between these categories such that they are equally activated, NAM predicts that some stimuli may not be categorized at all. Luce and Pisoni (1998) provide some evidence for this, finding a correlation between "frequency-weighted neighborhood probability"—that is, the probability of a stimulus being categorized as a particular word, given lexical competition from that word's neighbors and frequency effects—and accuracy of identification of words in noise. Moreover, in a lexical-decision task, the experimenters found that reaction times were higher and accuracy rates lower for nonwords with higher

[5] Because no one has ever explicitly tested for the FL, however, much of the experimental work to be discussed only has indirect implications for the FL.

neighborhood density and frequency—that is, lexical competition caused a decrease in accuracy and processing speed.

There is also evidence at the sub-lexical level that listeners are sensitive to sub-phonemic variation. McMurray et al. (2003), for example, conducted an eye-tracking experiment in a visual world paradigm to look at listener sensitivity to within-category VOT changes. Participants heard stimuli manipulated on a 9-point VOT scale, ranging from unambiguously [pa] to unambiguously [ba]. They had "pa" and "ba" on the screen and had to click on the button that best represented the stimulus they just heard. Unsurprisingly, listeners' responses were categorical and highly consistent. However, items that had more ambiguous VOTs resulted in higher activation of the other phoneme, as reflected by fixation times. This suggests that listeners are sensitive to sub-phonemic variation. In a second experiment, McMurray et al. (2003) pictured minimal pairs differing in word-initial voicing (e.g. *bear* and *pear*) in a visual world paradigm experiment. Again, more ambiguous VOTs resulted in higher activation for the competitor word. These were similar results to those in McMurray et al. (2008), in which sub-phonemic VOT manipulations resulted in relatively long-lasting differences in competitor activation for "lexical garden-paths," like *barricade* vs. *parakeet*. These results suggest

that listeners are sensitive to sub-phonemic differences, which is ultimately required by the FL. The results also reinforce those of Luce and Pisoni (1989), since competition caused by sub-phonemic differences manifests itself in processing.

Perhaps the strongest evidence for the FL comes from Nielsen (2011), in an imitation experiment. Participants were exposed to words beginning with /p/'s whose VOTs had been artificially elongated by 40ms. Beforehand they recorded a baseline token, and, afterwards, a post-exposure token. Participants consistently imitated, as they produced significantly longer VOTs in the post-exposure recordings. Moreover, this effect generalizes to the sub-lexical level, since similar increases in VOT were observed for novel /p/-initial words and even novel /k/-initial words. Frequency was also negatively correlated with imitation. Exemplar theoretic approaches predict this, since higher frequency words have more exemplars and more strongly entrenched categories. Each new exemplar, then, makes up a smaller ratio of the total number of exemplars, and thus has a smaller effect on future productions.

Crucially, in a second experiment, Nielsen (2011) found no imitation whatsoever when VOTs of target words were *reduced* by 40ms. If all

exemplars were created equal, so to speak, we might expect both types of manipulated stimuli—long VOT and short VOT—to have an equal effect on the productions of speakers. The FL, however, predicts some of the short—and therefore more ambiguous—VOT stimuli would not be stored, thus leading to reduced imitation. This is reflected in Nielsen's results.

Krajlic and Samuels (2006), on the other hand, found that stimuli containing stops ambiguous for voicing could engender perceptual learning. A phoneme ambiguous between /t/ and /d/ was inserted in contexts where it could only be /t/ ("fron**t**ier") or /d/ ("han**d**y") to create the critical words. The same manipulated phoneme was inserted in both contexts. There was also a control group: the critical words in this case also contained the ambiguous phoneme, but were nonwords. Participants were exposed to these stimuli in a lexical decision task, after which they identified consonants in VCV form, which had been manipulated on a six-point VOT continuum.

The researchers found an effect of perceptual learning for the experimental group, and none for the control group, suggesting that ambiguous phonemes have more effect on listeners when they are encoded as a part of a lexical item. This supports the idea of an interface between

the lexical and phonemic levels, per Pierrehumbert (2002). Similar effects were found in Norris, McQueen, and Cutler (2003) and Kraljic and Samuel (2005). In addition, Vroomen et al. (2007) found these effects when the disambiguating factor was visual: participants were exposed to a nonword ambiguous between /aba/ and /ada/, which was paired with video of a speaker producing either /b/ or /d/. Those participants who saw visual /b/ more often showed the results of perceptual learning through an identification task.

While these results show that speakers are capable of being influenced by ambiguous stimuli, it does not necessarily contradict the predictions of the FL for a number of reasons. First, the FL does not suggest that *all* ambiguous tokens are filtered out—in Wedel's (2006) model, even perfectly ambiguous exemplars only have a 10% chance of being dropped, which, over the course of thousands of iterations of a perception/production loop, is enough of a selective pressure against ambiguity to prevent merger. Thus some perceptual learning is not surprising. For this to have contradicted the FL, the researchers would have had to have shown that stimuli manipulated for longer-than-normal VOTs—and are thus unambiguous—result in equal or less perceptual learning. Second, in the case of Kraljic and Samuels (2006), while these

segments were ambiguous at the phonemic level, they did not result in ambiguity at the lexical level. Vroomen et al. (2007), on the other hand, used nonwords. The present study differs in this regard, since target stimuli consisted of stop-initial minimal pairs ambiguous for voicing, resulting in lexical ambiguity.

In addition, Vroomen et al. (2007) found that participants exposed to unambiguous stimuli showed greater signs of perceptual learning. More importantly, as the number of exposures increase for ambiguous stimuli, perceptual learning decreases. After 256 exposures, the effect is completely gone, while it increases monotonically for unambiguous stimuli. Similar results—a diminishing effect of ambiguous stimuli over time—were found in Samuel (2001). Vroomen and Baart (2009) expand upon this, testing perceptual learning the day after exposure; no effect whatsoever was found. That said, the time-course of these effects is controversial, with other studies (Norris et al., 2003; Eisner and McQueen, 2006) finding effects lasting hours.

If, however, we take Vroomen et al.'s (2007) results at face value, this adds an interesting wrinkle to our story. A crucial feature of many exemplar theoretic models (originally proposed by Nosofsky, 1986; see also

Pierrehumbert, 2001; 2002; Wedel, 2006) is the exponential decay of exemplars over time, giving recent episodes a greater influence. The results of Vroomen et al. (2007) might suggest that the selective pressure against ambiguous tokens (i.e. the FL) isn't manifested at categorization, since significant perceptual learning initially took place. Instead, perhaps this bias takes the form of ambiguous tokens decaying more rapidly. That said, there may not be a straightforward way to model this. In addition, the FL seems less stipulative, as it is consonant with models like NAM and can be explained as an effect of lexical competition.

4. Dispersion effects

4.1 Universal typological trends

Universal trends within phoneme inventories has received more attention than any other possible dispersion effect (e.g. Lindblom and Liljencrants, 1972; Lindblom, 1986; Sanders and Padgett, 2008; Becker-Kristal, 2010; Flemming, 1995; 2004; Padgett, 2003; Boersma and Hamann, 2008; etc.). While there are exceptions to these trends, the number of inventories that follow them is striking. Each of these trends is consistent with an exemplar-based model of dispersion utilizing the Filtering Listener (see section 2.4 for discussion). What follows here is a straightforward description of the trends commonly cited as dispersion effects.

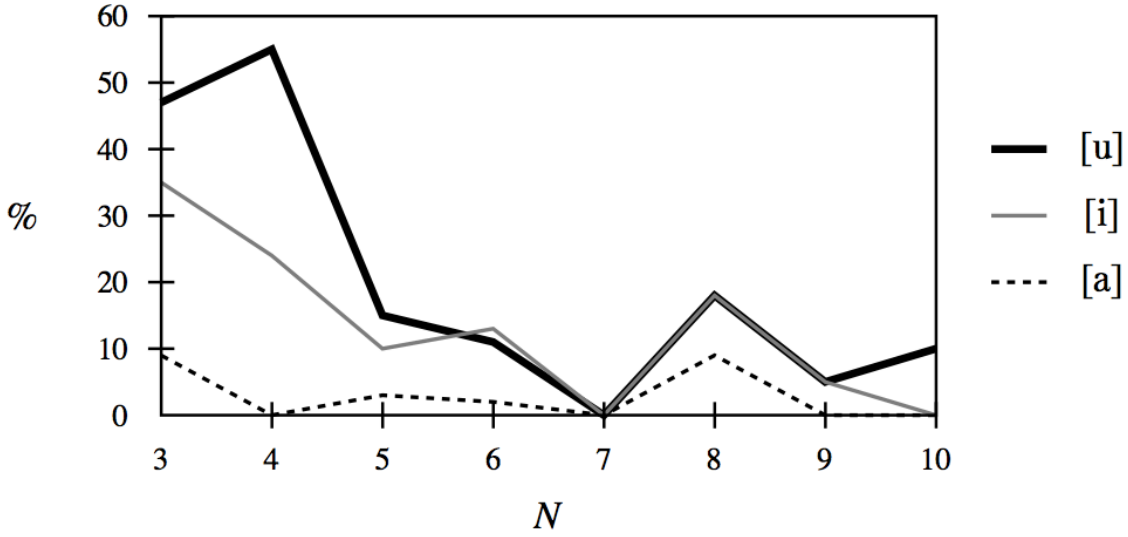
As Boersma and Hamann, 2008, note, languages with a single high vowel tend to have the high central vowel [ɨ], like Kabardian (Choi, 1989; 1991) and Marshallese (Choi, 1995). Speaking informally, we might say that with only one high vowel there is no need to use the full front-back color spectrum. If we assume an articulatory pressure to centralize (i.e. that more dispersed vowels have a higher articulatory cost), then, all things being equal, we predict inventories with one high vowel to prefer [ɨ], the

high vowel with the lowest articulatory cost.[6] [7] Inventories with two vowels, on the other hand, tend to avoid [ɨ] in favor of [i] and [u], presumably since these corner vowels offer a higher perceptual distance. Finally, inventories with three high vowels tend to have [i], [ɨ], and [u], making the most of the available perceptual space. We might then characterize the trend informally as the following: vowels centralize as much as possible, possibly to reduce articulatory cost, while keeping the minimum perceptual distance between each other required for maintaining contrast. (See Lindblom, 1990b, for more formal discussion).

[6] Bybee (2004) objects to the strong version of this hypothesis—that *all* sound change is due to reduction of articulatory effort—based on changes due to perceptual confusability, for example, or common fortition, epenthesis, and lengthening processes. That said, adopting a weaker version of this hypothesis—that there is a pressure to reduce articulatory effort that can effect *some* sound change—strikes me as uncontroversial.

[7] Flemming, 1995, points out that this constitutes a markedness reversal in a theory that posits universal markedness constraints like *ɨ >> *i, *u. This suggests that the concept of markedness should probably be considered relative to the inventory at hand. See Padgett, 2003, for discussion.

Figure 4: Percentage of N-vowel systems in UPSID missing corner vowels. From Sanders and Padgett (2008)



The figure above, reproduced from Sanders and Padgett (2008), shows the percentage of languages in the UPSID database (Maddieson, 1984; Maddieson and Precoda, 1989) missing one of the three corner vowels. Very few systems with five or more vowels lack one of the three corner vowels, but a considerable number of smaller systems lack [u] or [i]. We might infer that the high corner vowels seem to have an articulatory cost that is avoided if possible. Once systems reach a certain size, however, the acoustic distance between vowels is small enough such that the need for perceptual contrast is greater than the articulatory cost of hyper-articulated vowels.

Becker-Kristal (2010), using a corpus of acoustic descriptions of phoneme inventories in 230 languages, similarly finds systematic trends towards maintenance of perceptual distance.

Figure 5: Reproduced from Becker-Kristal (2010)

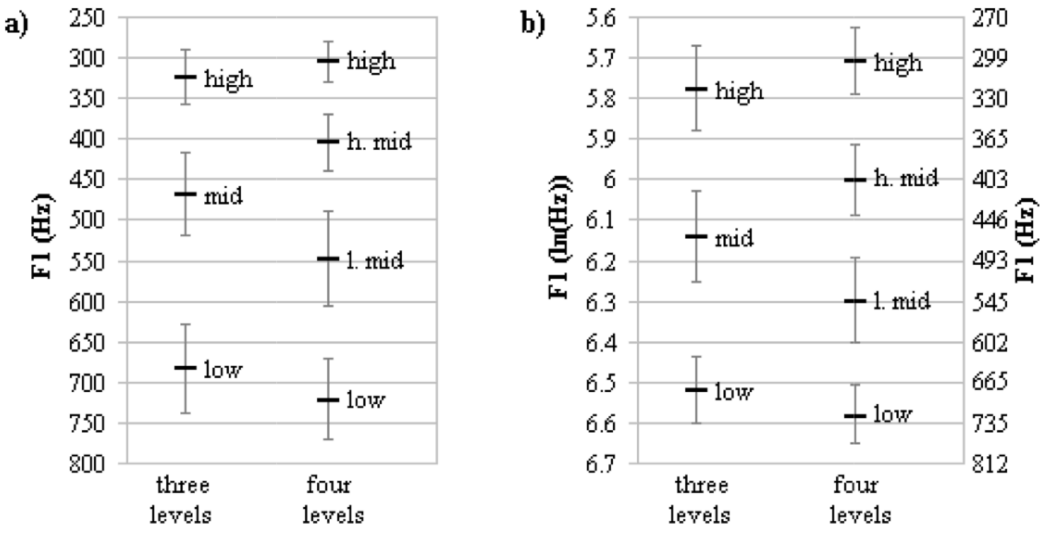


Figure 4-9: Averages and standard deviations of characteristic F1 frequencies of vowel height levels in peripheries with 5-6 vowels (three heights) and with 7-8 vowels (four heights), on the Hz scale (a) and the ln(Hz) scale (b). Hz equivalents of the ln(Hz) values are shown on the right edge of the right panel.

Here we see vowel height (F1) as a function of the number of peripheral vowels in the system. Thus having four, rather than three, peripheral height levels will, in the aggregate, push the high and low vowels further towards hyper-articulation. The result is a larger overall F1 range, creating more spacing between vowels, ostensibly to maintain sufficient perceptual distance. Note the remarkably even F1 spacing between vowels,

especially in the figure on the right using ln(hz) scale. Becker-Kristal (2010) also finds an effect of inventory configuration on the phonetic realization of a language's low vowel: in languages with "right-crowded" vowel inventories, the low vowel was significantly fronted relative to that in symmetric inventories. "Left-crowded" systems, on the other hand, result in significantly backed low vowels.

So far, we have seen four of the dispersion effects in Boersma and Hamman's (2008) typology: "preference for the centre," or an articulatory pressure towards centralization, seen in single high vowels surfacing as [ɨ]; "the excluded centre" in systems with an even number of vowels along a given dimension, such as languages with two high vowels, [i] and [u]; "the growing space", in which systems with more vowels use more of the available acoustic space; and "equal auditory distances". Boersma and Hamman (2008) also cite "permissible variation" as a dispersion effect: a phoneme will have a wider variety of phonetic realizations (i.e. allophones) "if it is alone on its auditory continuum than if it has neighbours from which it has to stay distinct" (Boersma and Hamman, 2008: 222). Again, we might suspect this is a result of the need for sufficient perceptual distance—if a phoneme is "crowded" by other phonemes on some given dimension, it is less likely to have allophones

that might overlap with the phonetic realization of those crowding phonemes. Boersma and Hamman (2008) note that languages with a single labial voiceless plosive often have voiced allophones (e.g. Central Arrernte: Breen & Dobson, 2005); languages with single high central vowels commonly have [i] and [u] as allophones (e.g. Kabardian: Choi, 1991); and that "languages with smaller vowel inventories show more vowel-to-vowel coarticulation" (Boersma and Hamman, 2008: 22) (e.g. Bantu langauges: Manuel, 1990). It should be noted that the evidence presented here is mostly impressionistic; a more exhaustive typological study is necessary before we can confidently characterize this as a dispersion effect.

4.2 Ambiguity in speech

Ambiguity is inherent in the speech signal along a number of dimensions. The production of a phoneme is affected by a large number of factors specific to the speaker and phonetic context, to the point that the production of different phonemes often overlap. This "lack of invariance" has been a significant avenue of research in speech perception, and has caused considerable difficulties for machine speech recognition (e.g. Bernstein & Franco, 1996) even though human listeners handle it fluidly (Creelman, 1957; Verbrugge, Strange, Shankweiler, & Edman, 1976). As Newman, Close, and Burnham (2001) note, this variability "can be caused

by...dialect (Byrd, 1992), social group (Johnson and Beckman, 1997), speaking rate (Miller and Liberman, 1979), emotional state (Shankweiler, Strange, and Verbrugge, 1977), gender (Byrd, 1992), vocal tract length (Fant, 1973; Peterson and Barney, 1952), articulatory habits (Johnson and Beckman, 1997; Klatt, 1986), and phonetic context (Liberman et al., 1967)."

In the following figure from Hillenbrand, et al. (1995), an update on Peterson and Barney's (1952) classic study of the American vowel space, we can see the vowels of 140 men, women, and children (with each plotted vowel representing one speaker's mean formant values for productions of that vowel).

Figure 6: The vowel space of American English speakers. Reproduced from Hillenbrand et al. (1995). 46 men, 48 women, and 46 children are pictured.

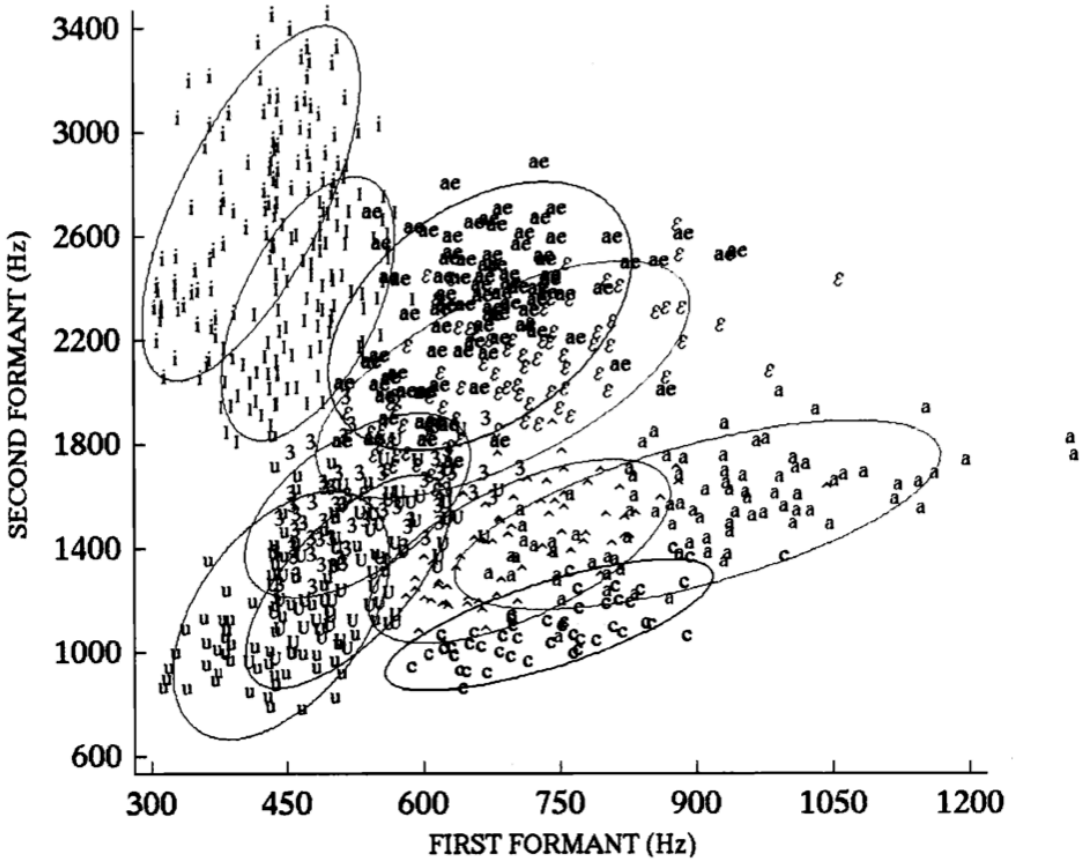


FIG. 4. Values of $F1$ and $F2$ for 46 men, 48 women, and 46 children for 10 vowels with ellipses fit to the data ("ae"=/æ/, "a"=/ɑ/, "c"=/ɔ/, "ʌ"=/ʌ/, "ɜ"=/ɝ/). Measurements for /e/ and /o/ have been omitted, and the data have been thinned of redundant data points.

There is quite a large degree of overlap, especially for the more central vowels. While much of the variance here is due to the differences between men, women, and children, inter-speaker differences between speakers of the same gender and dialect have been shown to be significant (e.g. inter-speaker production differences in Johnson, Ladefoged, and Lindau, 1993;

VOT differences in Theodore, Miller, DeSteno, 2009; etc.).

Among all this ambiguity, how are speakers able to keep phonemic categories separate? This is all the more vexing in light of evidence that speakers' productions are influenced by the percepts they encounter. This is borne out by experiments in Goldinger (1998) and Goldinger (2000), as well as the growing body of literature on imitation (e.g. Babel, 2011; Namy, Nygaard, and Sauertig, 2002; Pardo, 2006; 2009; among others). Unsurprisingly, speakers' productions can also be influenced over time by dialects they are exposed to (e.g. Harrington, 2000).

These studies strongly suggest that exemplar storage is multi-layered—one exemplar might include the acoustic characteristics of a primary phonemic or lexical category (e.g. the vowel /a/), as well as tags for individual speaker and various socio-indexical markers. Thus exemplars from a familiar speaker will be highly activated upon hearing a new percept from that speaker; this might explain the "voice effect" found in Goldinger's studies (see Johnson, 1997 for discussion of weighted activation for individual speakers). Additionally, we might expect a child's speech, for example, to highly activate those exemplars stored from previous examples of children's speech. So even if there is phonetic

overlap between an adult and child speaker for some phoneme, the fact that exemplars tagged with "child" are highly activated will ensure the correct categorization.

This provides a possible exemplar theoretic mechanism for listeners' ability to parse phonetically ambiguous speech—that is, normalization. Still, while speaker and socio-indexical tags will highly activate certain exemplars, this does not mean that exemplars without those tags will be completely excluded. While Johnson (1997) assumes complete exclusion in his model for simplicity's sake, he notes, "obviously, real listeners do not absolutely rule out impermissible alternatives" (Johnson, 1997: 150). This mechanism may also be inadequate for when listeners encounter new voices, for which they have no distribution of exemplars. Additionally, if we assume complete exclusion, we might predict that speakers' productions might only be influenced by voices of speakers matching in gender, age, socio-economic class, and ethnicity. While speakers may be influenced *more* by those within similar societal groups, not being influenced at all by members of other groups seems unlikely for a number of reasons. For example, listeners have been found to imitate speakers of the opposite sex (e.g. Goldinger, 2000).

So despite the normalization mechanism of exemplar theory, we still expect significant ambiguity in the speech stream due to inter-speaker variation. In addition, we might predict ambiguity due to within-speaker variation as well. While there is much less research on this topic, Newman et al. (2001) found significant within-speaker differences in the productions of /s/ and /ʃ/ of English speakers, with some speakers exhibiting more variability than others. The researchers also conducted an identification task using the fricative stimuli produced, and found that listeners' reaction times were longer for speakers who were more variable in their productions.

All of the ambiguity and variation inherent in speech poses a serious problem for contrast maintenance; the FL provides a possible mechanism militating against this ambiguity.

4.3 Maintenence of phonemic contrast and anti-homophony effects

Functionalist arguments for sound change have been made most notably by Martinet (1952, 1964, 1974), and were continued by Lindblom (1986) to explain the universal patterns found in section 2.1. Functionalists contend that there is a causal link between the maintenance of phonemic contrast and sound change. A merger of two phonemes that distinguish a large number of words (i.e. a high functional load) is less likely than that of two

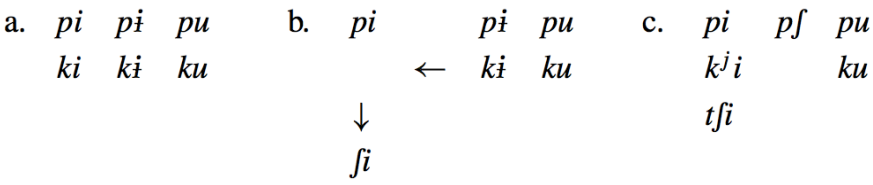
phonemes that would result in few homophones. The common objection to functionalism is that it is teleological—it assumes speakers are perhaps aware of collapsing phonemic distinctions and go out of their way to avoid it (see Blevins, 2004; Labov, 1994; Ohala, 1984). A compelling functional argument has been made, however, for some types of sound change.[8]

Padgett (2003), working within the framework of Dispersion Theory, explores the phenomenon of post-velar fronting in Russian from a functionalist standpoint. In East Slavic—a progenitor of modern Russian—velar consonants did not appear before front vowels, and could only precede the high central vowel [ɨ]. This differed from consonants at other places of articulation, which could appear before either [i] or [ɨ]. (This also meant non-velars were contrastive for palatalization, since there is an allophonic alternation in Russian that [i] follows palatalized consonants and [ɨ] follows non-palatalized consonants). The "gap" in the phonemic inventory—i.e. [pʲ] and [tʲ] but no [kʲ]—was left behind by another, earlier sound change in which "palatilized velars had mutated to

[8] Wedel, Kaplan, and Jackson (under revision) provide evidence in a corpus study that the number of phoneme mergers is negatively correlated with functional load (that is, the number of minimal pairs differentiated by a given contrast). This finding was extended in Wedel, Jackson, and Kaplan (under revision), which shows that similarity between minimal pairs in syntactic category and frequency inhibit merger. Together, these studies provide some evidence for a surprisingly simple version of the classic functional load hypothesis.

palato-alveolars (e.g. *kʲ* -> tʃʲ)". Then, between the 12th and 14th centuries, post-velar fronting occurred, where *kɨ* -> *ki*. (There is another, separate rule, palatalizing velars before front vowels).

Figure 7: Velarization in East Slavic. Reproduced from Padgett (2003).



What was the impetus for this change? It is difficult to explain as a result of articulatory phonetic pressures—it is certainly not an assimilation, and dissimilation seems unlikely as well (see Padgett, 2003, for discussion). The reason for the change might be functional: because of the "gap" that *kʲ* -> tʃʲ left in the inventory, the change would not result in a neutralization, which is not the case for consonants at other places of articulation. Here is a diagram of the three stages, reproduced from Padgett (2003):

This may explain why only velars underwent the change. Furthermore, the perceptual distance between the resulting *kʲ*i and ku is greater than

that between kɨ and ku—thus preserving and even enlarging the contrast.

The dropping of French coda /s/ in the plural article may present a possible case of compensatory sound change, initiated to avoid widespread homophony. As Labov (1994: 569) notes, the French article for plurals was once *las*, similar to Spanish. At some point the [s] was lost, although one suspects it still exists in the underlying representation, since it appears in liason when the following word begins in a vowel. Because most French nouns do not begin in vowels, however, this important contrast that arises very frequently would be lost in most cases. Possibly as a result, the contrast has manifested itself in differences in vowel quality between the feminine singular and plural determiners, with [la pɔm] for *the apple*, and [le pɔm] for *the apples*. The shift of contrast is not complete, as the number distinction is lost with the use of certain prepositions, such as *au* ("to the" (sg.)) and *aux* ("to the" (pl.)), both of which are produced as [o].

Vowel chain shifts present another example of a possible dispersion effect. A simple chain shift occurs when vowel A moves, and vowel B moves into the position previously occupied by vowel A:

Figure 8:

/B/ -> /A/ ->

If vowel A initiates the change, this is a *pull chain*. If vowel B initiates it, it is a *push chain*. Of course, chain shifts are often more complicated, with multiple vowels taking part. If vowel A had not moved in Figure 8, it would have resulted in a merger. Thus chain shifts are often given, at least in part, a functional explanation.[9] Note that a purely functional explanation is incomplete, since there are universal trends observed in chain shifts that seem to have little to do with preserving contrast (e.g. long vowels rise; short vowels fall; etc.). Interestingly, while there are cases that violate these trends, they are much more likely to occur when the violation results in a contrast being maintained. Labov (1994:270) writes, "The study of chain shifting does not then provide many examples of unexpected mergers; examples of unexpected avoidance are easier to locate"; such examples include diphthongal movements in southern

[9] Blevins (2004: 285) argues that functional explanations for chain shifts are problematic, since mergers do in fact occur. While this is true, it may be a case of the exception proving the rule, since merger avoidance is the much more common pattern. Moreover, there is no reason to suspect that merger avoidance would be absolute, given the complex set of possible causes for chain shifting (Labov, 1994: 328). Wedel, Kaplan, and Jackson (under revision) also provide strong preliminary evidence that functional load, in the form of minimal pairs, plays a role in contrast maintenance. See footnote 8 for more.

Swedish and parallel vowel shifting in Romansh.

Exemplar theory offers a straightforward and non-teleological mechanism to model chain shifts. We are less concerned with the cause of the initiating change, as much as the chain that follows. As the initiating change takes place, the distribution of speakers' exemplars along some acoustic dimension will shift. Using a highly idealized and simplified model, let's take the example of /i/ raising in the Great Vowel Shift, a pull chain. As /i/ raises, listeners store and produce exemplars with continually lowering F1 values, leaving a gap in the vowel space. The soon-to-be-pulled vowel now has an asymmetry on what it borders in the exemplar space. The tail of the distribution of exemplars for /e/ on the higher end of the F1 scale is bordered closely by /æ/, and, given the above evidence on variability in speech, we might expect the two categories to overlap to some degree. Exemplars in this ambiguous area are at a selective disadvantage for one of two reasons: either they are split between the two categories (the weak hypothesis) or sometimes simply not stored at all (the FL). On the other tail of the distribution for /e/—the low end of the F1 scale—there is no competing vowel. While there are, by definition, still few exemplars there, any instances of /e/ perceived within that acoustic range are stored without competition. This gives an advantage to

exemplars on the low end of the F1 scale. This selective pressure snowballs over time, with the entire distribution eventually shifting towards the gap left behind by /i/. This, of course, creates another gap, enabling shifts of multiple vowels. The same mechanism is employed in a push shift, simply with the inverse results. Note that in a push shift, the initiating sound change has to be powerful enough to overcome selective pressures against ambiguous exemplars. The pushed vowel, due to those same pressures, eventually moves away, setting off its own chain reaction.[10]

Finally, there are cases of anti-homophony, in which "an otherwise regular sound change can be locally inhibited where it would eliminate a crucial contrast" (Wedel, 2006: 20). Blevins (2005) argues that anti-gemination, such as in Iraqi Arabic, is such an effect, in that it is a phonetically unnatural process preserving a paradigmatic contrast that would otherwise be lost. Blevins and Wedel (2009) expand on this, citing, for example, unstressed vowel syncope in Dakelh. This is a case in which a phonetically natural process—vowel syncope—that exists in most forms of the language is avoided where it would erase a systematic contrast. In

[10] See Ettlinger (2007)—which posits two idealized speakers exchanging exemplars, rather than a single-seeded perception/production loop—for a more formal discussion and model implementation. See Blevins (2004: 288) for a similar account.

Proto-Athabaskan, the progenitor of Dakelh, the valence prefixes were *də, *ɬ, and *ɬə. Following historical syncope (Krauss, 1969, as cited in Blevins and Wedel (2009)), these forms changed to [d-], [ɬ-], and [l-]. There is a synchronic epenthesis process, however, which targets precisely those forms that were historically *ɬə (1st-person singular and 2nd-person plural). Otherwise, the historical *ɬə- and *ɬ-valence forms would be homophonous. Thus an otherwise regular sound change is inhibited to prevent homophony. While this follows the general pattern of contrast preservation we've seen, it differs in category type: this is an example of morphemic, rather than lexical or phonemic, categories avoiding merger. Of course, we could view morphemic contrast preservation as a subset of lexical preservation (since morphemes are always part of a word). But this seems indirect, overly complex, and may make the wrong predictions—if this contrast-avoidance occurred on a word-by-word basis, we would not expect to see such a systematic patterns, and would likely get numerous exceptions. This speaks to a mostly unresolved tension in exemplar theory over the "grain size" over category types—see Pierrehumbert (2002) for further discussion.

5. Present study

5.1 Introduction

The present study called for stimuli that were "skewed" ambiguous. If an ambiguous percept is defined as something that will be variably categorized—i.e. not consistently heard as the same word–then a perfectly ambiguous percept would be heard half the time as one member of a minimal pair, and half the time as the other member. A skewed ambiguous percept, then, might be categorized three-quarters of the time as one member of a minimal pair, and a quarter of the time as the other member. Using something that is skewed ambiguous is necessary for two reasons: first, a perfectly ambiguous stimulus has no "right" answer, which makes measuring accuracy, at least in any straightforward way, impossible.[11] Second, something somewhat, rather than perfectly, ambiguous should mitigate floor effects—a real concern since the stimuli are presented in noise, and thus are already difficult to comprehend.

A pilot study was run to define a baseline for what constitutes a "skewed

[11] If only one segment is manipulated for ambiguity, as is the case in this study (i.e. word-initial voicing) and stimuli are perfectly ambiguous, then accuracy scores can be tabulated based on the rest of the word. It is not entirely clear what the Filtering Listener predicts in terms of this measure.

ambiguous" stimulus. This was necessary for the main experiment, in order to understand how the stimuli presented to participants should be manipulated. Finding exact and exhaustive accuracy scores for stimuli with differing VOT values was not, however, the point of the pilot. This is because in the main experiment, accuracy scores were independently taken for each experimental block, and these were the crucial numbers. In other words, the purpose of the pilot was only to provide guidelines for stimulus creation for the main experiment.

A note about stimulus creation: word-initial stop voicing was chosen as the element to be manipulated for its simplicity. The main cue for stop voicing in English is VOT, something far simpler to manipulate than formant values of vowels, for example. Initially, stimuli were created by incrementally cutting out portions of aspiration of the voiceless member of a minimal pair to create the voiced version, but this resulted in unnatural-sounding stimuli, with listeners often reporting [h] or a vowel as the first segment.

It is possible that while cutting aspiration from the voiceless member of a pair results in a cue suggesting a voiced stop—that is, a short VOT—other cues were disrupted. In particular, a large portion of the formant

transitions—all transitions in the aspiration—for these manipulated "voiced" stops were cut out, resulting in unnatural-sounding stimuli. A different strategy was utilized: aspiration was added to the voiced member of the pair, i.e. VOT was lengthened. In particular, replacing the beginning of periodicity (i.e. the vowel) with aspiration will lengthen the VOT while also correspondingly modifying the formant transitions.[12] The result is that after manipulation both cues are consistent with a voiceless stop and the formant transitions are kept intact, resulting in more natural stimuli.[13]

5.2 Pilot 1

Participants

Three volunteers took the pilot, none of whom were familiar with linguistics. Participants were between the ages of 24 and 26, and were monolingual, native English speakers.

Procedure

The experiment was designed and presented in Praat (Boersma and Weenink, 2013). Participants were presented with a forced-choice task.

---

[12] Thanks to Janet Pierrehumbert for pointing this out.
[13] It may also be possible to avoid problems with the formant values by replacing aspiration from the voiceless member of the pair with a portion of periodicity from its voiced counterpart—in other words, just the opposite of what was done here.

Two boxes were presented on the screen, each with one member of a minimal pair differing in word-initial voicing (e.g. *pun*/*bun*). 500ms later the stimulus was played. Participants were instructed to select the word on the screen that sounded closest to what they heard. This was done with a keyboard, with participants pressing [f] for the word on the left and [j] for the word on the right. Participants had unlimited time to answer, although they were instructed to do so quickly.

Word list

10 sample minimal pairs were chosen from the list of 32 minimal pairs that would be manipulated for ambiguity in the main experiment. Each of these was stop-initial and monosyllalbic, with the stop constituting the only segment in the onset (i.e. words like *trap* were excluded). Of these 10, three of the word-initial stops were bilabial, three were velar, and four were alveolar. It has been shown that within speakers, place of articulation is the most important factor for length of VOT in English (Lisker & Abramson, 1964). Among the sample pairs, within each stop there was at least one low, one mid, and one high vowel, since vowel height has also been shown to be a significant factor in length of VOT. For each of the 10 sample words, a five-point scale, ranging from unambiguously voiced to unambiguously voiceless, was created, for a total

of 50 stimuli. Each stimulus was heard 4 times, for a total of 200 tokens.

Stimulus creation

The experimenter's voice was used for all recordings.[14] Using Praat, the first two periods from the vowel of the voiced member of the pair was removed and replaced with aspiration, equal in duration, from the voiceless member. The left and right boundary of the portion of the vowel that was being removed were always at zero-crossings. When copying aspiration from the voiceless counterpart, the left boundary of the aspiration was placed roughly 5ms following the end of the burst. This created one stimulus; using the same method, four more stimuli were created by removing 3, 4, 5, and 6 periods.

Results

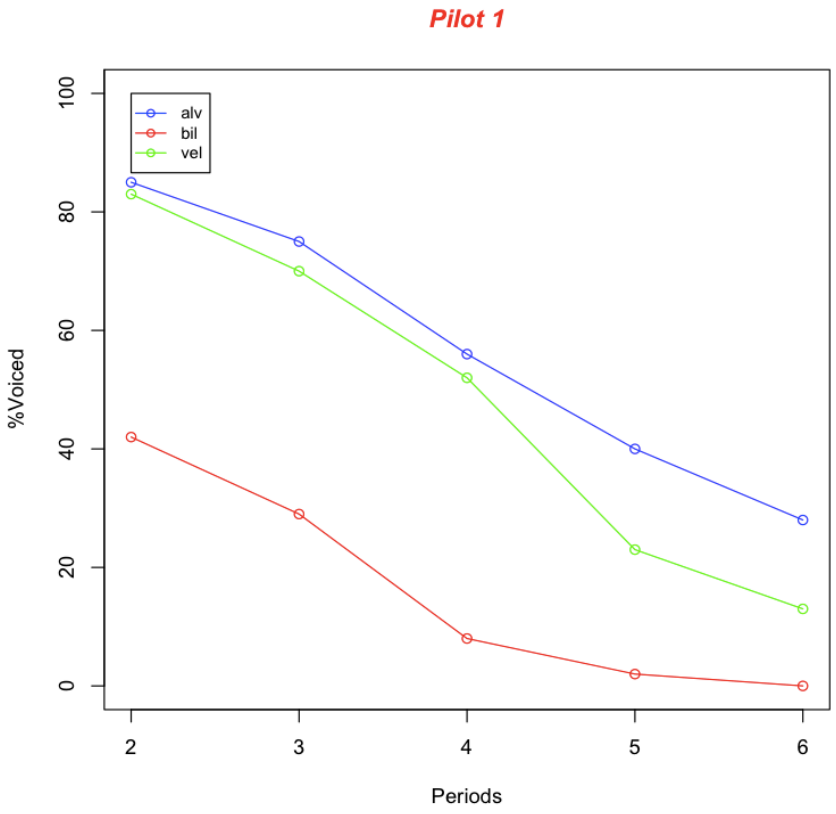
On the x-axis of Figure 9 are the five stimuli, each with a different number of periods removed. The percentage of the time that participants responded to hearing those stimuli as the voiced member of the pair is on the y-axis. As expected, hearing a stimulus as voiceless correlates with the

[14] While experimenter bias is always a factor to consider, it did not seem important in this case since (a) the experimenter had no conscious control over his VOT times, (b) there was no particular result that was more favorable than another, and (c) these VOT times were manipulated, in any case.

number of periods that are removed. This relationship is close to being linear for velar and alveolar stops; bilabial stops, overall, are more likely to be heard as voiceless. This is not surprising, since bilabial stops have the shortest VOTs (CN). We would thus expect them to be affected more categorically by manipulation, and may need a more fine-grained approach. Moreover, removing only two periods still results in a voiced rate under 50%, suggesting that a different scale needs to be presented to the listener that includes fewer periods removed. Thus another pilot was conducted.

Vowel height was not a significant factor for VOT.

Figure 9: First segment voicing discrimination by place in Pilot 1.



5.3 Pilot 2

Participants

Two UCSC linguistics graduate students participated, both in their early twenties and native speakers of English (one was a bilingual Spanish speaker).

Procedure

See Pilot 1.

Word list

The same three bilabial pairs from Pilot 1 were used. Only one velar and one alveolar from Pilot 1 were used; they were included to refine the results from Pilot 1. Each pair was arranged on a five-point scale, leading to 25 stimuli. Each of these was heard eight times, resulting in 200 total tokens.

Stimulus creation

The method of stimulus creation was identical to Pilot 1, but the scale changed, ranging from one to three periods removed. The scale was, however, in half period increments, again resulting in a five-point scale: 1, 1.5, 2, 2.5, and 3 periods removed. While removing half a period is unnatural, it seemed that such a tiny difference would be below the level of consciousness for the listener. These periods ranged from 7ms to 9ms; it was hypothesized that replacing portions of the vowel as small as 3.5-4.5ms with aspiration would not result in unnatural stimuli or be particularly problematic, even though "half periods" in and of themselves are highly unnatural. Again, the left and right boundaries of all portions of the vowel replaced with aspiration were at zero-crossings, even when half a period was replaced.

Results

As can be seen in Figure 10, the results for bilabials for Pilot 2 are indeed more fine-grained and less categorical than those in Pilot 1. That said, finding stimuli that are heard as voiced roughly 75% of the time is difficult; the bilabial stimuli with only one period removed were heard as voiced only 47% of the time.

Discussion

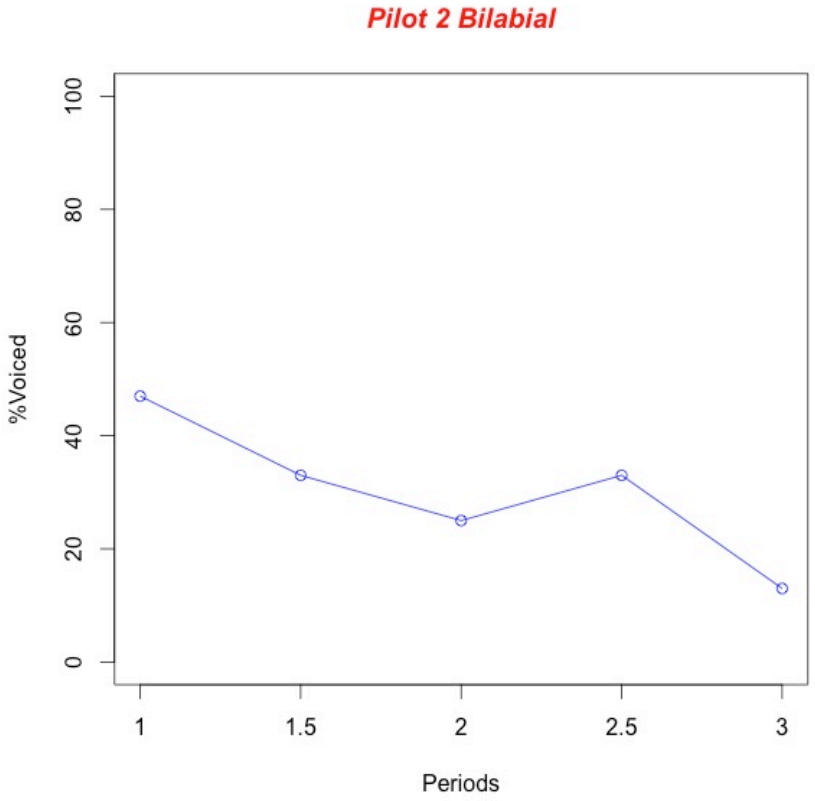
The number of periods that were ultimately replaced for the main experiment can be seen in Table 1. The percentage of tokens heard as voiced for that number of periods in the two pilot studies is in parentheses.[15]

Table 1: Number of periods replaced by aspiration, by place. Percent heard as voiced in parentheses.

	Bilabial	Alveolar	Velar
Voiced	1 (47%)	3 (75%)	2 (83%)
Voiceless	2 (25%)	6 (28%)	5 (23%)

[15] The values for velars and alveolars are taken from Pilot 1; bilabial values are from Pilot 2.

Figure 10: First segment voicing discrimination in Pilot 2 (bilabial only).



5.4 Methodology

Word list

The word list was made up of 72 minimal pairs, the members of which were monosyllabic and stop-initial, differing only in first-segment voicing. There were 36 ambiguous target pairs (that is, pairs whose word-initial VOTs were manipulated) and 36 unambiguous control pairs (no manipulation). A third of the pairs began with an alveolar stop, a third velar, and a third bilabial. The entire list is given in the appendix.

Frequency, neighborhood density, and neighborhood frequency were all considered as possible confounds when compiling the word list. Differences in these measures may have had an effect in two areas: between members of a minimal pair and between the set of ambiguous words and the set of unambiguous words.

There is very robust evidence that frequency affects processing. Luce and Pisoni (1989), for example, found that participants were more accurate in perceptual identification and auditory lexical decision tasks for high-frequency words. As another example, Dahan et al. (2001) monitored eye movements in a task in which participants followed spoken instructions to

move pictures on a monitor with a mouse. The experimenters found that participants fixated more often on high-frequency words than low-frequency words. In a second experiment in which target words had no competitors, participants fixated more quickly on high-frequency words. Frequency has also been shown to affect sound change (Bybee, 2001; Pierrehumbert, 2002; c.f. Dinkin, 2008), with high frequency words more likely to undergo lenition and vowel centralization.

Neighborhood density and neighborhood frequency have also been shown to affect processing. Luce and Pisoni (1989) found that both words with low-frequency neighbors and words in sparse neighborhoods tend to be recognized more quickly than those with high-frequency neighbors and dense neighborhoods. Moreover, both density and frequency have been shown to affect production—in particular, low-frequency and high-density words tend to have more dispersed vowels (see Wright, 2004; Munson and Solomon, 2004; Munson 2007; among others).

As noted above, there were two areas in which these effects could lead to a confound, one being the differences between members of a minimal pair. The effects of neighborhood density are unlikely to have a large effect here, since the word list was made up of minimal pairs, and differences in

neighborhood density are usually quite small between members of minimal pairs (since they share so many neighbors).

Differences in frequency between members of a minimal pair, however, can be quite large. Thus members of pairs were matched for frequency, although this was not always possible for a large list of monosyllabic words differing only in word-initial voicing—a set of constraints that resulted in a narrow set of possible words to draw on. All frequency data was taken from the $SUBTLEX_{US}$ corpus (Brysbaert and New, 2009), which is made up of subtitles from American film and television.

Table 2 shows the log frequency of the word list broken down by ambiguity and voicing. At first glance differences do not appear to be significant, and an ANOVA confirms this: differences in log frequency based on ambiguity, voicing, and place of articulation were not statistically significant.

Table 2: Log frequency of word list.

	Ambiguous	Unambiguous	Voiced	Voiceless	Total
Min.	1.40	0.90	0.90	1.40	0.90
Median	2.85	2.87	2.90	2.84	2.85
Mean	2.89	3.03	2.99	2.93	2.96
Max.	5.12	6.06	5.50	6.06	6.06
SD	0.74	1.10	1.02	0.85	0.94

Participants

Participants consisted of 28 UC Santa Cruz undergraduate students, all of whom received class credit. All but 3 participants were native English speakers—their results were not included in the d' analysis or LME model.

Stimulus creation

Stimuli were created using the same method as in the pilot study. For each of the 36 ambiguous target minimal pairs, both the voiced and voiceless members of the pair were manipulated to create "skewed ambiguous" (SA) stimuli. As described above, periods from the onset of the vowel were removed from the voiced member of the pair and replaced with aspiration from the word-initial stop of the voiceless member of the pair. The number of periods replaced with aspiration depended on the stop and whether it was an SA voiced or voiceless stop (see Table 1).

Pink noise was created through Praat with Niels Reinholt Petersen's "Create Waveforms" script. The distribution was Gaussian; all other settings were left as the default.

Procedure

Participants were tested in groups of three or fewer in a sound-attenuated booth. The experiment was designed and run in E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Participants were given instructions that they would be listening to words, one at a time, in background noise, and should type the word that they hear. A red crosshair was fixated in the center of the screen throughout the experiment. At the onset of the trial, pink noise, scaled to 75db using Praat's "Scale intensity" feature, was played. 500ms later, the stimulus, scaled to 70db, was played (the pink noise continued to play over the stimulus). [16]

After the stimulus and pink noise ended, a cursor appeared on the screen. Participants could then type their response, which they could see as they typed. Participants could use backspace to delete typing errors, and had unlimited time to answer, although they were instructed to answer quickly. After the response had been typed, participants pressed enter to

[16] In Goldinger (1996), a warning phrase is used, after which white noise is played. 50ms later, the stimulus is played. In the present study the onset of noise, by coming a full 500ms before the stimulus, also acted as the warning that a stimulus was about to be played. This was done to avoid particularly disrupting perception of the crucial manipulation, which was made to the initial segment of the stimulus. A very short (e.g. 50ms) period of noise played before the stimulus—especially because noise and aspiration share some acoustic qualities—might skew perception of the voicing of the first segment.

continue to the next trial. After a 2-second interval, the next trial began.

In each block, participants were presented with randomly ordered words from the 72 minimal pairs, half of which were ambiguous and half unambiguous. There were two versions of the experiment differing in which member of the minimal pair participants heard. In other words, in one version of the experiment, a participant would be presented with "core", while in the other version another participant would be presented with "gore". Thus every participant was presented with one member of each pair, but no participant was presented with both members of a pair. In both versions of the experiment, half the stimuli had voiced initial segments, and half had voiceless initial segments.

There were four identical blocks in the experiment (although order of stimuli was random for each one). In between blocks participants were allowed to take a short break.

5.5 Results and analysis

5.5.1 *Data analysis*

Two measures of accuracy were taken: that of the first segment alone, and that of the whole word. Both are relevant, as the "grain size" of stored representations—that is, lexical or phonemic—is very much at issue.

For the whole word measure, all homophones for each word form were counted as correct answers. Moreover, the data were corrected for common misspellings (e.g. "deen" for "dean") and phonetic spellings (e.g. "gool" instead of "ghoul"), both of which were counted as correct answers. Single-letter deletions or substitutions were also made (e.g. "betg" for "beg"), but only if the incorrect letter key was adjacent to the key for the correct letter. This method was only applied to nonwords. The first quarter of the data were analyzed for these types of misspellings; any identical mistakes in the rest of the data were also corrected. Less than 1% of the data were corrected for misspellings.

There were 28 subjects, each of whom completed 4 blocks with 72 trials, for a total of 8064 data points.

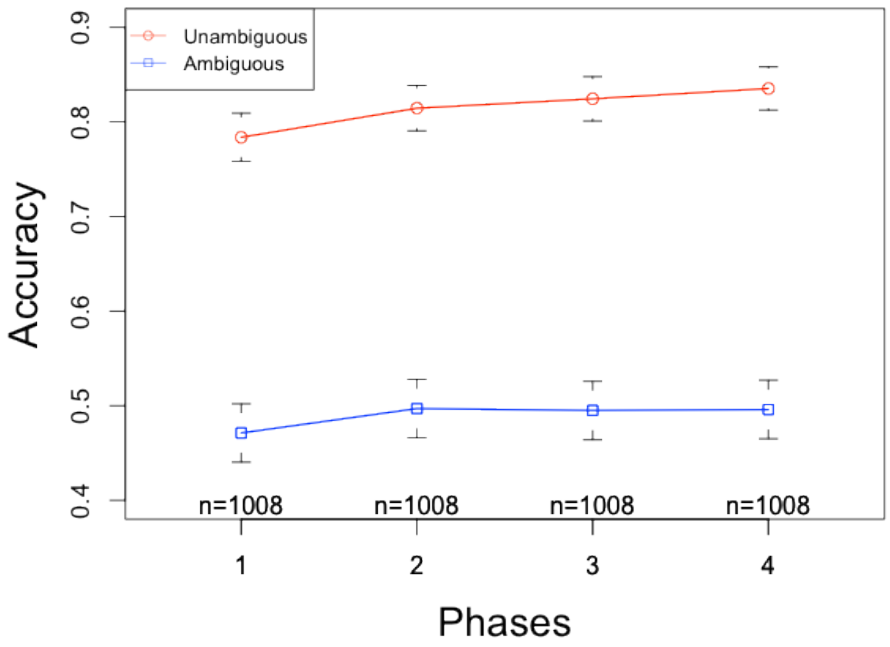
5.5.2 *Accuracy results*

Table 3 shows the overall accuracy rates for all subjects and items for entire words. Unambiguous stimuli improve over twice as much as ambiguous stimuli, providing some preliminary support for the FL. The same results are plotted in figure 11.

Table 3: Overall word accuracy by ambiguity and phase. Data from all stimuli and participants included.

	1	2	3	4	Total	Improvement
Ambiguous	0.47	0.50	0.50	0.50	0.49	0.02
Unambiguous	0.78	0.81	0.82	0.84	0.81	0.05

Figure 11: Accuracy by ambiguity over the course of the experiment. Data from all stimuli and participants included.



In Table 4, we can see the overall first-segment accuracy. Again, improvement is higher for the unambiguous condition, although the improvement is quite small, probably because participants were close to ceiling at over 90% accuracy. Unfortunately this makes comparing the two conditions quite difficult, and after excluding data from non-native speakers and aberrant stimuli (more on that soon), unambiguous stimuli are even closer to being at ceiling. Thus the analysis that follows is based on entire-word accuracy only unless otherwise noted, since the crucial

comparison can't be made between ambiguous and unambiguous conditions.

Table 4: Overall first-segment accuracy by phase and ambiguity. Data from all stimuli and participants included.

	1	2	3	4	Total	Improvement
Ambiguous	0.55	0.55	0.55	0.56	0.55	0.01
Unambiguous	0.90	0.91	0.92	0.92	0.91	0.02

Unfortunately, participants had a large bias towards hearing stimuli as voiced, as can be seen table 5. In fact, they reported hearing the first segment of a stimulus as voiced 18% more often than voiceless.

Table 5: First segment response type. Data from all stimuli and participants included.

	Response %
Voiceless	0.34
Voiced	0.56
Other	0.10

The confusion matrix in table 6 sheds light on exactly what type of errors participants made.

Table 6: Confusion matrix, with each cell containing a percentage of error type. This table includes all incorrect responses in the data. Columns organized by type of stimulus, and rows by type of error. SA = skewed ambiguous; rhyme = nucleus and coda. "Onset voicing/other/no onset and rhyme" type errors are responses with multiple errors. Error categories are mutually exclusive; thus "Onset voicing only" is not a subset of "Onset voicing and rhyme", for example.

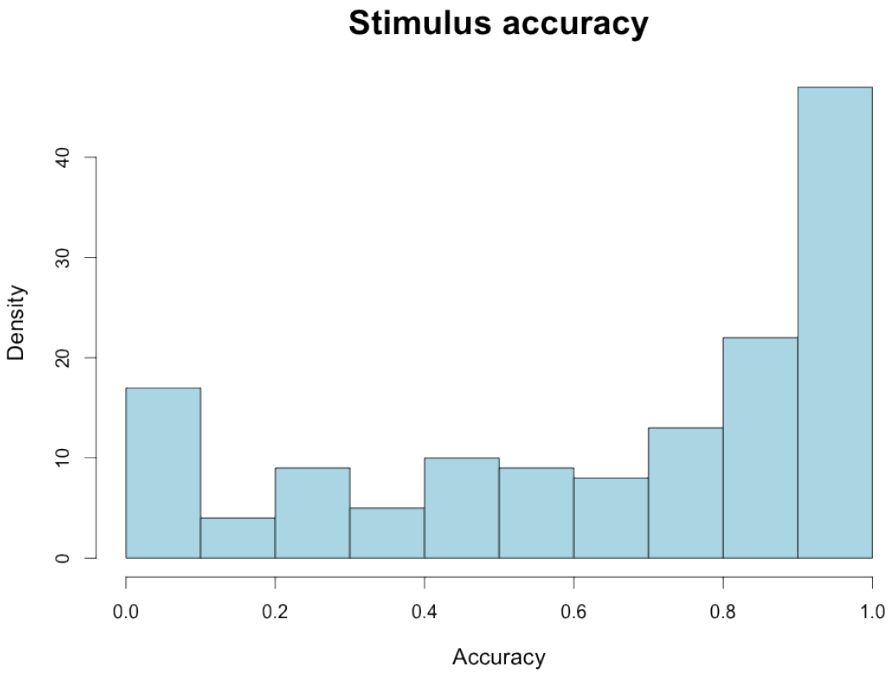
	Unambiguous voiced	Unambiguous voiceless	SA voiced	SA voiceless	Total
Onset voicing only	1%	0%	3%	24%	28%
Onset other only	1%	4%	1%	2%	9%
No onset only	1%	0%	1%	1%	3%
Rhyme only	8%	7%	7%	5%	27%
Onset voicing and rhyme	1%	0%	4%	14%	19%
Onset other and rhyme	1%	2%	1%	1%	4%
No onset and rhyme	0%	0%	3%	6%	9%
Total	**13%**	**13%**	**21%**	**53%**	**100%**

The majority of errors came from ambiguous voiceless stimuli. Note that a large proportion of errors (27%) occur only in the rhyme, and thus have nothing to do with word-initial voicing. In fact, these make up the majority of errors in the unambiguous condition, reflecting the results shown in table 4 in which participants were close to ceiling in the unambiguous condition for first-segment voicing. Note also the difference in error rates between voiced and voiceless stimuli in the ambiguous

condition. They are similar in terms of rhyme-only errors, but onset-only errors are much more numerous for ambiguous voiceless stimuli (24%) than voiced (3%). This reflects participants' general bias towards voiced responses, as shown in table 5.

This bias was so strong that a number of SA voiceless stimuli were *always* heard as voiced. In Figure 12, we see a histogram of accuracy scores by stimulus, binned by 10%.

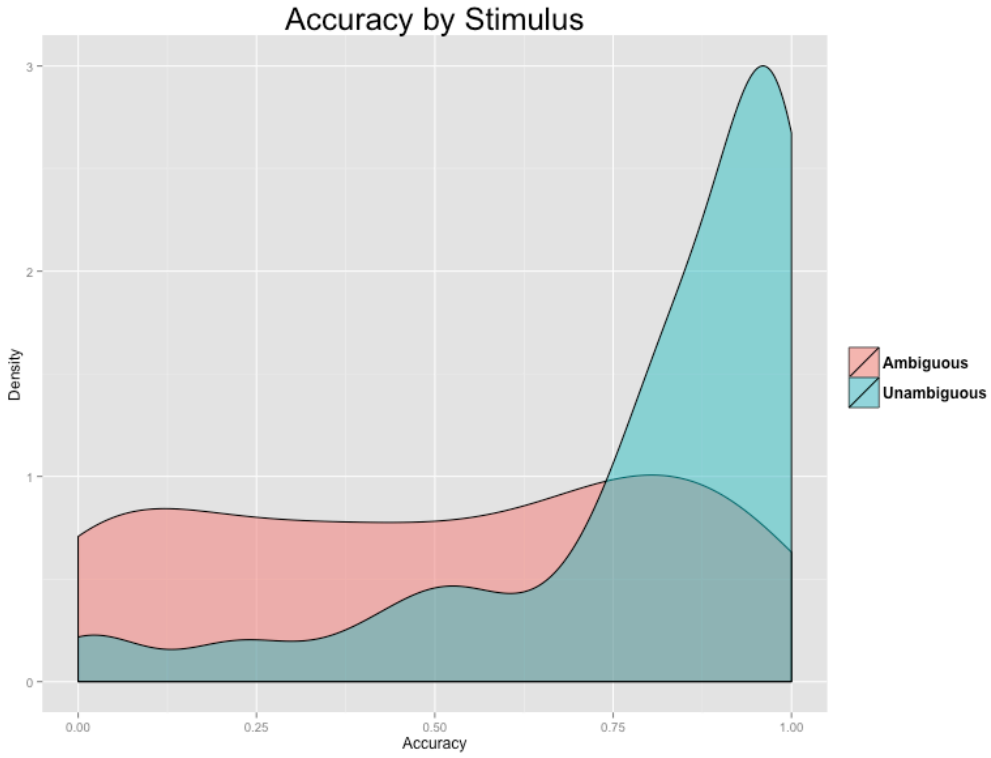
Figure 12: Histogram of whole-word accuracy by stimulus. Binned by 10%. Data from all stimuli and participants included.



There are a large number of stimuli (13) whose accuracy rates are below 10%. As can be seen in Figure 13, most of these low-accuracy stimuli were in the ambiguous condition (10/13). While these 10 stimuli were intended to be skewed ambiguous, they were, in actuality, totally unambiguous: virtually every participant heard the opposite (i.e. voiced for voiceless) of what the experimenter intended. These stimuli will be excluded from the data in the analyses in the following sections (i.e. for d' and LME models), on the grounds that they were flawed, and, as a result, almost universally misinterpreted. This accounted for 12% of the total data.[17] (Note that Figure 13 is very similar to Figure 12, only broken up by ambiguity and not binned discretely).

[17] It should be noted that, because these stimuli had such low overall accuracy, they showed little improvement over the course of the experiment. In addition, most of them were part of the ambiguous condition. These data, although not very meaningful, would have helped support the FL hypothesis, which predicts more improvement in the unambiguous condition than its ambiguous counterpart. If removing them affects the results, it should only make them more conservative.

Figure 13: Gradient histogram of whole-word accuracy by stimulus, broken up into ambiguous and unambiguous conditions. Data from all stimuli and participants included.



Furthermore, all but one of the low-accuracy stimuli were voiceless, reflecting the overall bias of participants towards voiced responses. This bias was exclusively limited to the ambiguous stimuli (Figure 14), suggesting that it was not task-specific, but a function of the way in which the stimuli were manipulated.

Figure 14: Whole-word accuracy by voicing and ambiguity. Data from all stimuli and participants included.

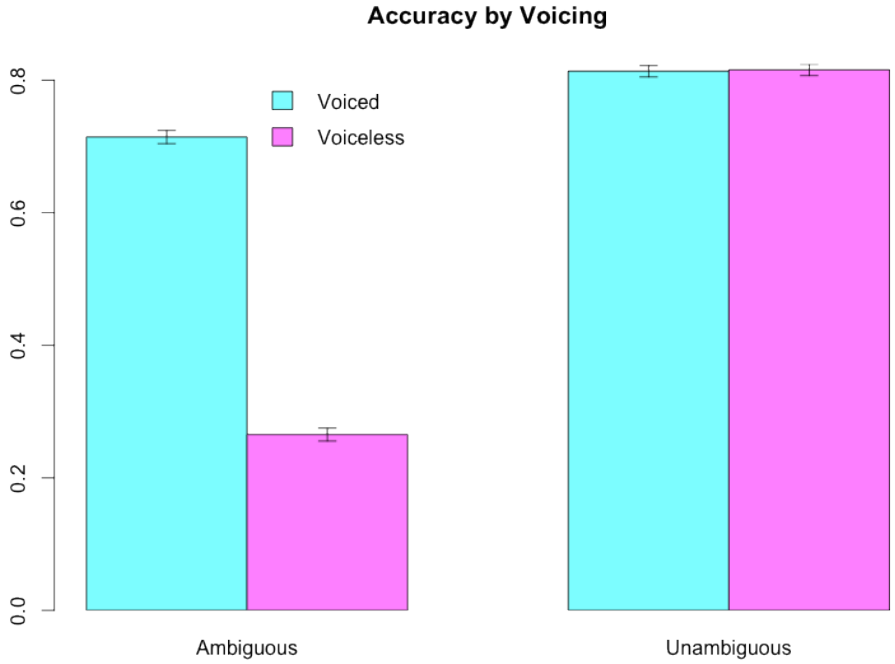
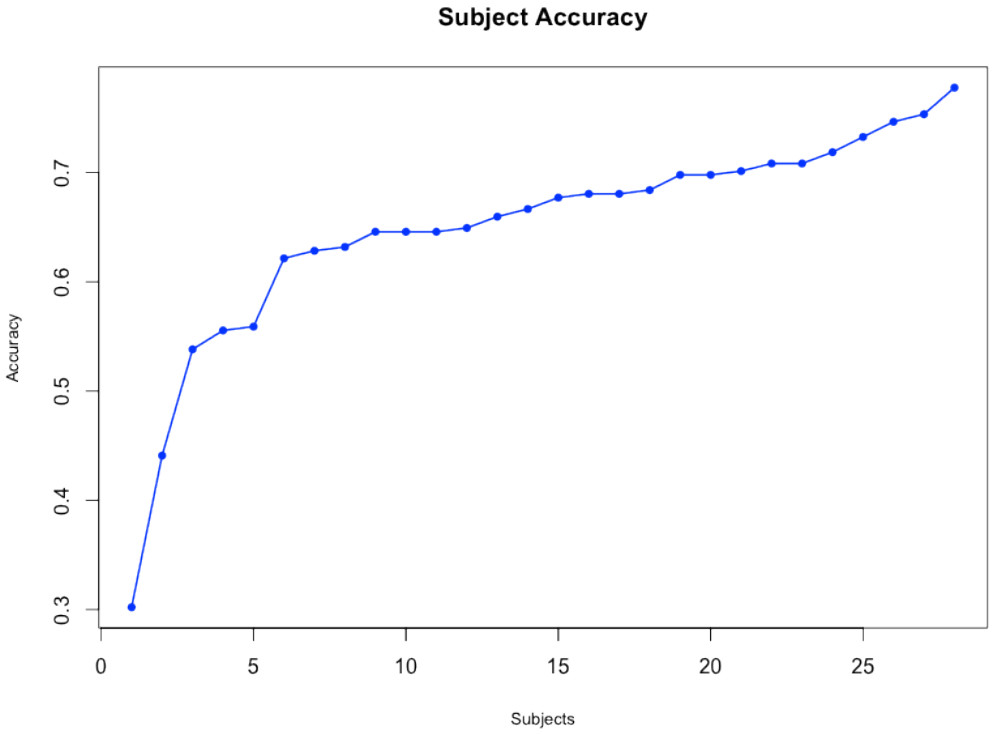


Figure 15 shows accuracy by subject. All 28 subjects are shown here; in the following sections, the 3 non-native speakers were excluded, for a total of 25 speakers. Non-native was defined as a speaker who learned English older than four-years-old. Unsurprisingly, these 3 speakers had the 3 lowest accuracy rates.

Figure 15: Whole-word accuracy by subject. Data from all stimuli and participants included.

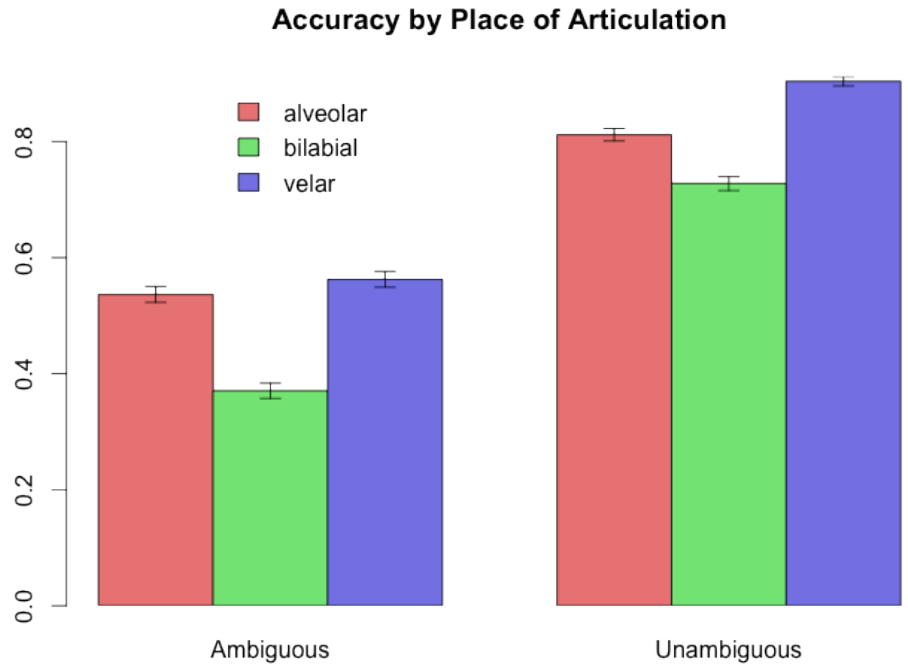


Finally, accuracy by first-segment place of articulation—for all subjects and stimuli—can be seen in Figure 16. Broadly speaking, alveolar and velar stops are both fairly accurate relative to bilabial stops. Looking more closely, 8.5% of participants' responses began with something other than a stop—mostly (80%) this consisted of either a vowel or an *h*. Virtually all (97%) responses beginning with an *h* (e.g. *hat* given as a response for the stimulus *bat*) came from stimuli beginning with a bilabial stop. Moreover,

about two-thirds of responses beginning with a vowel were from bilabial stimuli.

This is not totally surprising, given that bilabial stops have the shortest VOT of the three places of articulation, and will thus be the most sensitive to manipulation. It follows that if VOTs were shortened too much, participants might not hear a word-initial stop at all.

Figure 16: Whole-word accuracy by place of articulation of stimulus word-initial segment. Data from all stimuli and participants included.



5.5.3 *d' analysis*

d' is widely considered a more informative measure than raw accuracy, since it gives you overall sensitivity. As noted earlier, in this and the following sections 3 speakers were excluded, as well as the 16 least accurate stimuli. In addition, all of the "neither" responses—that is, responses that did not begin with a stop—were also excluded, since we are primarily interested here in ambiguity over voicing. That leaves 5700 data points, or 71% of the original data set.

While d' is usually used for discrimination tasks, it is easily adapted to the current experiment. Because there is no "same" and "different" here, we arbitrarily labeled voiced responses as "same" and voiceless as "different"; the results would be the same if the opposite were the case. d' measures sensitivity by looking at hits (i.e. voiced things that were correctly categorized as voiced) and false alarms (voiceless things that were incorrectly labeled voiced). A d' analysis also allows us to look at how biased participants were towards giving one type of answer over another (i.e. voiced or voiceless).

Table 7: Whole-word d' by phase and ambiguity. Change in accuracy from previous phase is in parentheses. Total improvement is the change in accuracy from Phase 1 to Phase 4.

	Ambiguous	Unambiguous	Total
Phase 1	0.72	2.74	1.53
Phase 2	0.78 (.06)	2.90 (.26)	1.62
Phase 3	0.73 (-.05)	3.02 (.12)	1.62
Phase 4	0.72 (-.01)	3.15 (.13)	1.66
All phases	0.73	2.95	1.60
Total improvement	0	.41	.13
Bias	-0.53	0.07	-0.32

The ambiguous condition has much lower scores overall, which is to be expected. We also see a large bias (negative scores mean a bias towards voiced responses) for ambiguous stimuli, but not for unambiguous stimuli, reflecting what we saw in Figure 15. Most importantly, the difference in improvement between the two conditions is stark, with the ambiguous condition seeing virtually no improvement whatsoever, and the unambiguous condition improving significantly.

An ANOVA was run to establish the significance of these observations. Because subjects were repeating a single task and giving multiple observations per condition, they were treated as random effects. Thus this was a repeated measures ANOVA, with phase and ambiguity as factors.

Table 8: d' repeated measures ANOVA, with subjects as random effects.

	Df	Sum sq	Mean sq	F value	Pr(>F)	Significance
Error: Subject	24	19.400	0.808	-	-	
Error: Subject:Phase	3	0.452	0.151	1.461	0.2324	
Residuals	72	7.432	0.103	-	-	
Error: Subject:Ambiguity	1	211.216	211.216	326.378	1.78E-15	***
Residuals	24	15.532	0.647	-	-	
Error: Subject:Phase: Ambiguity	3	0.718	0.239	2.802	0.0459	*
Residuals	72	6.152	0.085	-	-	

The ANOVA reveals that the overall difference in d' over the course of the experiment (i.e. by phase) was not significant. Difference in d' due to ambiguity was highly significant, which is to be expected. Crucially, there is a significant interaction between phase and ambiguity, suggesting that improvement over the course of the experiment is affected by ambiguity. This is predicted by the FL hypothesis.

5.5.4 *Linear mixed-effects models*

A linear mixed-effects model was implemented to try and tease apart any effects caused by ambiguity, which are relevant to this experiment, from those caused by the voicing bias in the stimuli. I will discuss two models

here, which differ in their results. The detailed results of both models can be found in the appendix.

The first model includes phase, ambiguity, and voicing as independent variables, with accuracy as the dependent variable. In addition, interactions between (1) phase and ambiguity, and (2) phase and voicing were added. Subjects were included as a random effect. The model was implemented using the lme4 package (Bates and Sarkar, 2007) in R (R core team, 2013).

Ambiguity ($F[1,2] = -11.2$, $p<.001$) and voicing ($F[1,2] = 3.6$, $p<.001$) were both significant. Voicing showed a significant interaction with phase 2 ($F[1,2] = 3.5$, $p<.001$), phase 3 ($F[1,2] = 3.8$, $p<.001$), and phase 4 ($F[1,2] = 5.0$, $p<.001$).[18] This suggests the voicing bias increased over the course of the experiment; bias measures from d' support this. Ambiguity, on the other hand, only showed a significant interaction with phase 3 ($F[1,2] = -2.0$, $p<.05$) and phase 4 ($F[1,2] = -2.9$, $p<.01$). This represents the crucial interaction, and confirms the predictions of the FL even when voicing, which presents its own highly significant effect, is also included in the model. It also suggests that ambiguity matters more as the course of the

[18] Note that each phase was measured against phase 1; differences in adjacent phases should also be measured in any future work.

experiment continues. This is expected, since changes by condition in accuracy from filtering out ambiguous stimuli over multiple exposures should be cumulative.

In the second model, the interaction between ambiguity and voicing was added. These factors show a highly significant interaction ($F[1,2] = -12$, $p<.001$), which is expected, given the fact that voicing only seemed to affect accuracy for ambiguous stimuli, as shown in Figure 15.

In this model, the crucial interaction of ambiguity and phase ceases to be significant. It follows a similar trend to the first model, however, in that it gets closer to significance as the experiment goes on: phase 2 ($p = 0.90$), phase 3 ($p = 0.39$), phase 4 ($p = 0.14$).

Also note that which of the two word lists participants were exposed to (i.e. if they heard *bat* and *tab* or *pat* and *dab*, for example) was not a significant predictor of whole-word accuracy or change in whole-word accuracy over the course of the experiment.

6. Discussion and Conclusion

6.1 Summary of results

The results of the experiment are promising, if not entirely conclusive. The crucial comparison is between the ambiguous and unambiguous condition, as the FL predicts different rates of improvement for each of them. We have seen that raw accuracy rates support this prediction, as do d' scores (for which the differences between the ambiguous and unambiguous conditions were found to be significant in a repeated measure ANOVA). Two linear mixed-effects models produced differing results, with only one finding a significant interaction between ambiguity and phase.

It is obvious that the manipulated stimuli were inherently flawed, as participants had a strong bias towards hearing ambiguous stimuli as having voiced onsets. Unfortunately this bias increased over the course of the experiment. This is particularly problematic because it only affected one condition, making comparing the ambiguous and unambiguous conditions difficult. Thus in the second LME model, when we added the interaction of voicing and ambiguity, the crucial effect was no longer

significant. It seems that changes over phase in accuracy are better predicted by voicing than ambiguity. Nevertheless, we still saw the interaction between ambiguity and phase moving closer to significance over the course of the experiment, although it never reached significance.

This voicing bias constitutes a formidable confound. Furthermore, participants were virtually at ceiling at distinguishing the first segment of unambiguous stimuli, which makes a true comparison between ambiguous and unambiguous conditions difficult for this measure. In turn, this experiment has limited things to say about grain size–that is, whether exemplars are stored at the lexical or phonemic level—since we could only measure storage at the lexical level.

6.2 Design issues

A number of simple design changes could be made to militate against some possible confounds. First, there was a separate word list for each of the unambiguous and ambiguous conditions. If the words on one of these lists were, for whatever reason, easier to improve upon, it would result in an important confound, since these were the crucial conditions. Differences in log frequency between the lists, which presents the most obvious possible confound, were not found to be significant in an ANOVA.

Still, any future experiment should draw upon one word list with the ambiguous words counter-balanced across participants.

Second, the voicing bias suggests that the pilot studies were inadequate, and a much more statistically powerful and tightly controlled pilot is required. In addition, the stimuli in the pilot studies were not heard in noise, as they were in the experiment. This results in two problems: first, participants unexpectedly turned out to be extremely good at judging the first segments of unambiguous stimuli, despite the noise. Second, any interaction between the addition of noise and the manipulation of the stimuli was not was not visible in the pilot studies. Such an interaction might have led to the voicing bias.

Perhaps Vroomen et al. (2007) offers a better method altogether for determining ambiguity. Before the main experiment, each participant completed a calibration phase, in which they completed a discrimination task, classifying stimuli on a /aba/ to /ada/ continuum. This enabled the experimenters to determine each participant's /b/-/d/ boundary—the most ambiguous stimuli was then used in the experiment for that participant. This allows for a more finely tuned calibration of individual listeners'

perceptual boundaries, and what counts as ambiguous for each participant.

6.3 Future work

Other than a more tightly controlled replication, there are a number of avenues for future research. One is a replication of Nielsen (2011), in which stimuli with lengthened VOT resulted in imitation, but those with shortened VOTs did not (see section 3.2 for a more thorough discussion). While Nielsen (2011) lengthened and shortened VOTs by the same amount (40ms) to create both types of stimuli, a more useful test of the FL would be to make sure the shortened-VOT stimuli were actually ambiguous for individual participants. This could be accomplished using the method described above, from Vroomen et al. (2007).

The role of context as a disambiguating factor is a crucial question, and one that is quite open. An ambiguous phoneme might be disambiguated by three contexts: a visual context, such as that explored in Vroomen et al. (2007); a highly predictable lexical context, like ***b****asketball*; and some higher level sentential or pragmatic context, which could disambiguate a segment that is not disambiguated by lexical context, like the minimal pairs found in this study (e.g. [b] or [p] in *?at*). One experiment might

address these three types of disambiguation, comparing how they engender perceptual learning.

The grain size of stored representations—a crucial issue for exemplar theory—is another open question, and one that has appeared repeatedly through this paper. An interesting question is how grain sizes and learning interact: does perceptual learning from individual segments, or segments embedded in nonwords, generalize to the lexical level, and vice versa? Certainly exemplar theory predicts the effect would be attenuated: in this framework, we have separate distributions of exemplars for lexical categories and each of the phonemic categories that make up those lexical categories. Thus perceptual learning from individual segments should not engender equal learning in lexical items, or else the latter would be nothing more than the sum of its parts, and would not require its own level of representation. On the other hand, there must be some interface between these two levels: our distribution of phonemic exemplars comes from lexical items (since we hear words, not individual segments, in real speech), and our production of neologisms, for example, must come from concatenated phonemic exemplars. Just how much each level affects the other, however, is not immediately clear.

6.4 Conclusion

We have seen dispersion effects proposed as a cause for a number of related phenomena: phonologization; compensatory sound change; anti-homophony; universal trends within vowel inventories; and chain shifts. All of these phenomena can be unified by the assumption that language has a mechanism for maintaining phonemic contrast. Most proposals up to this point explaining this mechanism have been speaker-based, and teleological in nature. More importantly, it has never been empirically shown that speaker-based effects can apply to the maintenance of specific phonemic contrasts, and are not just global in nature.

Listener-based, exemplar-theoretic approaches present two mechanisms for achieving dispersion effects. The weak hypothesis predicts that all exemplars are stored, but because ambiguous exemplars are split between multiple categories, they do not contribute to the distribution of a category as much as those in the mean of the distribution. It is not clear whether or not this mechanism is powerful enough—Wedel (2006) argues that there may be a system-wide reversion to the mean, caused by factors in perception or production. These system-wide pressures call for a stronger hypothesis: the Filtering Listener. The FL predicts that ambiguous words or phonemes are not always stored to phonetic memory. This increases the

selective advantage of unambiguous exemplars, leading to maintenance of categories in exemplar models despite any system-wide pressures.

This study was an attempt to test the predictions of the FL: that ambiguous stimuli are intrinsically degraded in memory. The experiment consisted of a perceptual identification task, in which participants undergo four identical blocks of identifying ambiguous and unambiguous stimuli in pink noise, following Goldinger (1996). Any improvement over the course of the experiment is assumed to reflect successful storage of exemplars. The null hypothesis is that ambiguity is not a significant predictor of improvement. All stimuli were made up of members of a monosyllabic minimal pairs that differed only in first-segment voicing. VOTs of first segments were manipulated to create the ambiguous stimuli.

The results were mixed, with analyses of accuracy and d' confirming the FL's predictions, but only one of two LME models showing ambiguity as a significant predictor of improvement. The results were muddied by two confounds: participants had a strong bias towards hearing ambiguous stimuli as voiced, and were extremely good at identifying the first segment of unambiguous words. A replication, with tighter controls and a

more thorough pilot study, is required for further confirmation of the FL's viability.

The FL has broad implications for sound change. Keeping phonemes distinct over time is absolutely crucial for phonology, and language as a whole, to function. The FL provides such a mechanism.

Appendix

Figure (i): Word list

Key: N (Minimal pair number); ND (Neighborhood Density); NF (Raw Neighborhood Frequency); LF (Log Frequency)

Unambiguous set						Ambiguous set					
N	**Word**	**Place**	**ND**	**NF**	**LF**	N	**Word**	**Place**	**ND**	**NF**	**LF**
1	tear	alveolar	36	23893	3.14	**37**	tab	alveolar	28	177	2.47
1	deer	alveolar	38	4143	2.65	**37**	dab	alveolar	27	947	1.76
2	torque	alveolar	17	1076	1.58	**38**	tile	alveolar	NA	4757	2.04
2	dork	alveolar	NA	NA	2.33	**38**	dial	alveolar	35	1305	2.66
3	tote	alveolar	36	522	1.75	**39**	tense	alveolar	19	382	2.72
3	dote	alveolar	29	618	1.08	**39**	dense	alveolar	18	381	2.04
4	tore	alveolar	41	18533	2.63	**40**	tent	alveolar	27	997	2.95
4	door	alveolar	40	17564	4.17	**40**	dent	alveolar	24	779	2.26
5	tug	alveolar	25	23087	2.15	**41**	toe	alveolar	45	64347	2.81
5	dug	alveolar	29	1692	2.66	**41**	dough	alveolar	44	34014	2.91
6	tied	alveolar	37	3176	3.14	**42**	tomb	alveolar	26	27972	2.46
6	died	alveolar	38	4303	3.9	**42**	doom	alveolar	33	7229	2.46
7	time	alveolar	28	798	5	**43**	tart	alveolar	19	980	2.09
7	dime	alveolar	26	2839	2.79	**43**	dart	alveolar	18	741	2
8	ten	alveolar	40	6627	3.87	**44**	tuck	alveolar	37	26896	2.61
8	den	alveolar	43	13486	2.5	**44**	duck	alveolar	41	2166	3.1
9	to	alveolar	31	91121	6.06	**45**	town	alveolar	20	2794	4.1
9	do	alveolar	47	74005	5.5	**45**	down	alveolar	28	7976	4.88
10	teen	alveolar	36	17302	2.32	**46**	tech	alveolar	31	5486	2.51
10	dean	alveolar	32	12545	3.3	**46**	deck	alveolar	36	1436	3.08
11	tool	alveolar	31	27570	2.74	**47**	tear	alveolar	34	10226	3.14
11	dual	alveolar	34	7074	1.76	**47**	dare	alveolar	36	9417	3.45
12	tune	alveolar	34	26284	2.9	**48**	tip	alveolar	35	23531	3.15
12	dune	alveolar	37	15095	1.72	**48**	dip	alveolar	37	3710	2.61
13	pill	bilabial	45	3912	2.78	**49**	pan	bilabial	45	10654	2.8
13	bill	bilabial	51	6407	3.78	**49**	ban	bilabial	47	14834	2.21
14	patch	bilabial	28	527	2.95	**50**	path	bilabial	21	343	3.1
14	batch	bilabial	28	3214	2.33	**50**	bath	bilabial	25	3144	3.2
15	peek	bilabial	34	1125	2.44	**51**	pun	bilabial	41	15940	1.98

15	beak	bilabial	34	8690	2.03	**51**	bun	bilabial	47	20359	2.17
16	pie	bilabial	41	53464	3.17	**52**	pore	bilabial	37	17599	1.4
16	bye	bilabial	39	57836	3.96	**52**	boar	bilabial	40	18031	2.05
17	palm	bilabial	10	129	2.83	**53**	pear	bilabial	34	8468	1.84
17	bomb	bilabial	36	1226	3.44	**53**	bear	bilabial	35	8934	3.47
18	park	bilabial	22	537	3.57	**54**	pin	bilabial	39	14311	2.92
18	bark	bilabial	26	455	2.45	**54**	bin	bilabial	42	15341	2.44
19	pad	bilabial	36	3385	2.62	**55**	pole	bilabial	35	1168	2.81
19	bad	bilabial	43	5037	4.44	**55**	bowl	bilabial	39	1513	3.04
20	pig	bilabial	26	1111	3.3	**56**	pat	bilabial	45	19112	2.97
20	big	bilabial	33	2709	4.54	**56**	bat	bilabial	48	25719	3.02
21	pond	bilabial	20	80	2.51	**57**	pump	bilabial	17	151	2.81
21	bond	bilabial	21	248	3.2	**57**	bump	bilabial	18	171	2.8
22	pot	bilabial	43	11345	3.06	**58**	punch	bilabial	10	187	3.18
22	bought	bilabial	44	15544	3.64	**58**	bunch	bilabial	15	1919	3.48
23	pack	bilabial	42	2914	3.35	**59**	pound	bilabial	21	1443	2.85
23	back	bilabial	47	1712	5.01	**59**	bound	bilabial	22	1540	2.97
24	peer	bilabial	35	1435	1.9	**60**	peg	bilabial	20	311	3.27
24	beer	bilabial	37	3949	3.59	**60**	beg	bilabial	25	1549	3.42
25	cod	velar	36	4301	2.06	**61**	card	velar	28	1078	3.64
25	god	velar	27	6117	4.66	**61**	guard	velar	19	1357	3.47
26	core	velar	39	20230	2.7	**62**	cap	velar	42	6797	2.98
26	gore	velar	35	17593	1.75	**62**	gap	velar	30	268	2.34
27	curl	velar	36	2248	2.08	**63**	came	velar	29	5161	4.37
27	girl	velar	23	307	4.45	**63**	game	velar	22	2046	4.08
28	cot	velar	46	11531	2.01	**64**	cane	velar	43	6658	2.63
28	got	velar	31	12740	5.23	**64**	gain	velar	40	2371	2.85
29	cash	velar	34	5614	3.57	**65**	cold	velar	29	2651	3.82
29	gash	velar	28	273	1.56	**65**	gold	velar	20	1990	3.6
30	come	velar	37	17589	5.2	**66**	coast	velar	23	647	3.13
30	gum	velar	30	14457	2.84	**66**	ghost	velar	16	543	3.27
31	cab	velar	30	5650	3.26	**67**	coup	velar	41	77290	2.13
31	gab	velar	26	302	1.56	**67**	goo	velar	38	81478	2.04
32	cage	velar	19	1023	3.01	**68**	calf	velar	24	5890	2.18
32	gauge	velar	17	754	2.05	**68**	gaff	velar	21	396	1.52
33	code	velar	38	2509	3.43	**69**	coat	velar	40	921	3.33
33	goad	velar	30	7943	0.9	**69**	goat	velar	37	12320	2.73
34	cave	velar	27	1394	2.85	**70**	could	velar	14	5926	4.92
34	gave	velar	23	1816	4.09	**70**	good	velar	15	8052	5.12
35	cool	velar	37	1978	4	**71**	coal	velar	43	2569	2.53

35	ghoul	velar	26	1030	1.72	**71**	goal	velar	36	5420	2.93
36	kilt	velar	24	977	1.51	**72**	cuts	velar	26	372	2.77
36	guilt	velar	18	4767	2.88	**72**	guts	velar	21	336	3.08

Table (i): Model 1 in section 5.5.4. Linear mixed effects model, with phase, ambiguity, and voicing as independent variables, and accuracy as the dependent variable. In addition, interactions between (1) phase and ambiguity, and (2) phase and voicing were added. Subjects were included as a random effect.

	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	0.9276	0.1166	7.9524	1.83E-15	***
Phase2	0.4100	0.1529	2.6818	0.00732	**
Phase3	0.3949	0.1530	2.5806	0.00986	**
Phase4	0.5184	0.1569	3.3042	0.00095	***
Unambiguous	1.7541	0.1539	11.3975	4.30E-30	***
Voiceless	-0.5439	0.1424	-3.8189	0.00013	***
Phase2:Unambiguous	0.2722	0.2270	1.1989	0.23055	
Phase3:Unambiguous	0.4640	0.2314	2.0049	0.04497	*
Phase4:Unambiguous	0.7117	0.2396	2.9708	0.00297	**
Phase2:Voiceless	-0.7309	0.2087	-3.5029	0.00046	***
Phase3:Voiceless	-0.8041	0.2099	-3.8305	0.00013	***
Phase4:Voiceless	-1.0803	0.2148	-5.0293	4.92E-07	***

Table (ii): Model 2 in section 5.5.4. Identical linear mixed effects model to that in table (i), with the addition of the interaction of ambiguity and voicing.

	Estimate	Std. Error	z value	Pr(>|z|)	Sig
(Intercept)	1.1995	0.1259	9.5271	1.62E-21	***
Phase2	0.4263	0.1625	2.6230	0.0087	**
Phase3	0.3854	0.1619	2.3802	0.0173	*
Phase4	0.4865	0.1652	2.9460	0.0032	**
Unambiguous	0.8590	0.1659	5.1788	2.23E-07	***
Voiceless	-1.1733	0.1547	-7.5853	3.32E-14	***
Phase2:Unambiguous	0.0261	0.2256	0.1156	0.9080	

Phase3:Unambiguous	0.1980	0.2293	0.8635	0.3878	
Phase4:Unambiguous	0.3487	0.2365	1.4744	0.1404	
Phase2:Voiceless	-0.6855	0.2111	-3.2482	0.0012	**
Phase3:Voiceless	-0.7097	0.2120	-3.3481	0.0008	***
Phase4:Voiceless	-0.9242	0.2155	-4.2881	1.80E-05	***
Unambiguous:Voiceless	2.0011	0.1674	11.9563	6.02E-33	***

References

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189.

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

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes (R package version 0.9975-11) [Computer software].

Becker-Kristal, R. (2010). Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus. Dissertation. UCLA.

Bernstein, J., & Franco, H. (1996). Speech recognition by computer. In N. J. Lass (Ed.), *Principles of Experimental Phonetics*, (pp. 408–434). St. Louis: Mosby.

Blevins, Juliette & Andrew Wedel (2009). Inhibited sound change: an evolutionary approach to lexical competition. *Diachronica 26*. 143–183.

Blevins, Juliette (2005). Understanding antigemination. In Zygmunt Frajzyngier, Adam Hodges & David S. Rood (eds.) *Linguistic diversity and language theories*. Amsterdam & Philadelphia: Benjamins. 203–234.

Blevins, J., & Wedel, A. (2009). Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica*, *26*(2), 143–183.

Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology*, *25*(02), 217.

Boersma, Paul & David Weenink. 2010. Praat: doing phonetics by computer (Version 5.1.45) [Computer program]. Retrieved October 28, 2012 from http://www.praat.org/.

Bradlow, A. R., Kraus, N., and Hayes, E. (2003). "Speaking clearly for children with learning disabilities: Sentence perception in noise," *J. Speech Lang. Hear. Res.* 46, 80–97.

Breen, Gavan & Veronica Dobson (2005). Central Arrernte. *Journal of the International Phonetic Association 35*. 249–254.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change*, *14*(03), 261-290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711-733.

Byrd, D. (1992). "Preliminary results on speaker-dependent variation in the TIMIT database," *J. Acoust. Soc. Am. 92*, 593–596.

Choi, John D. (1989). Phonetic evidence for a three-vowel system in Kabardian. *The Journal of the Acoustical Society of America, 86*. S18.

Choi, John D. (1991). An acoustic study of Kabardian vowels. Journal of the International Phonetic Association 21. 4–12.

Choi, John D. (1995). An acoustic-phonetic underspecification account of Marshallese vowel allophony. Journal of Phonetics 23.

Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America 29*, 655

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive psychology*, *42*(4), 317–67.

Dinkin, A. J. (2008). The real effect of word frequency on phonetic variation. *University of Pennsylvania Working Papers in Linguistics*, *14*(1), 8.

Ettlinger, M. (2007). An exemplar-based model of chain shifts. In *Proceedings of the 16th International Congress of the Phonetic Science* (pp. 685-688).

Fant, G.: Speech sounds and features (MIT Press, Cambridge 1973).

Fernald, A., and Kuhl, P. (1987). "Acoustic determinants of infant preference for motherese speech," *Infant Behav. Dev. 10*, 279–293.

Flemming, Edward. (1995). Vowels undergo consonant harmony. Paper presented at the Trilateral Phonology Weekend 5, University of California, Berkeley.

Flemming, E. (2004). Contrast and perceptual distinctiveness. *Phonetically-based phonology*, 232-276.

Garrett, A., & Johnson, K. (2011). Phonetic bias in sound change. *Origins of Sound Change: Approaches to Phonologization, Oxford: Oxford University Press, 53pp.*

Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. In *ISCA Tutorial and Research Workshop (ITRW) on Spoken Word Access Processes*.

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*(5), 1166–83.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251–79.

Guenther, Frank H., and Gjaja, Marin. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America 100*, 1111-1121.

Guy, Gregory R. (1996). 'Form and Function in Linguistic Variation', in G. R. Guy, C. Feagin, D. Schiffrin and J. Baugh (eds.), *Towards a Social Science of Language: Papers in Honor of William Labov, Volume 1: Variation and Change in Language and Society*, John Benjamins, Amsterdam, pp. 221–252.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000). Does the Queen speak the Queen's English? *Nature*, *408*(6815), 927–8.

Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, *130*(4), 2139–52.

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

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America, 97*, 553–562.

Johnson, K, Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*(2 Pt 1), 701–14.

Johnson, Keith, Edward Flemming, and Richard Wright 1993 The hyperspace effect: phonetic targets are hyperarticulated. *Language 69*, 505-528.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. *Talker variability in speech processing*, 145-165.

Johnson, Keith, & Mary E. Beckman (1997). Production and perception of individual speaking styles. In K. Ainsworth-Darnell & M. D'Imperio, eds., *Papers from the Linguistics Laboratory* (Ohio State University Working Papers in Linguistics 50), pp. 115-125.

Johnson, Keith. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology*, 25–40.

Kelso, J. A. Scott., Mingzhou Ding and Gregor Schöner. (1992). Dynamic pattern formation: a primer. In Arthur B. Baskin and Jay Mittenthal (eds.), *Principles of*

organization in organisms. Santa Fe Institute, Santa Fe, NM: Addison-Wesley Publishing Co., 397-439. 35

Kelso, J. A. Scott. (1995). *Dynamic patterns, The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Klatt, D. H. (1986). The problem of variability in speech recognition and in models of speech perception, in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt. Erlbaum, Hillsdale, NJ, pp. 300–324.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, *51*(2), 141–78.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, *13*(2), 262–8.

Labov, William Blackwell. (1994). *Principles of Language Change, Internal Factors*. Oxford: Blackwell.

Liberman, A.M., Harris, K.S., Hoffman, H.S., & Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*, 358–368.

Liljencrants, Johan & Lindblom, Bjorn. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. Lg 48. 839–862.

Lindblom, Björn. (1986). Economy of speech gestures. In Peter MacNeilage (ed.), The Production of Speech. New York: Springer-Verlag, 217-245.

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, The Netherlands), pp. 403–439.

Lindblom, B., & Maddieson, I. (1988). Phonetic Universals in Consonant Systems. *Language, Speech, and Mind*. 62-75

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

Luce, Paul. A., and David B. Pisoni. (1998). Recognizing spoken words, The neighborhood activation model. *Ear & Hearing, 19*, 1-36.

Martinet, André. (1955). Economie des changements phonetiques. Bern: Francke.

Martinet, André. (1964). Elements of General Linguistics, University of Chicago Press, Chicago.

Martinet, Andrée. (1974). Economía de los cambios phonéticos, Editorial Gredos, Madrid. [Translation into Spanish of Économie des changements phonétiques, 1955.]

McGuire, G., & Padgett, J. (2011). Explaining Dispersion Effects. *NYU Presentation*.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, *60*(1), 65–91.

McMurray, B., Tanenhaus, M. K., Aslin, R. N., & Spivey, M. J. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: evidence for gradient effects

of within-category VOT on lexical access. *Journal of psycholinguistic research*, *32*(1), 77–97.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel, 457–465

Munson, B. (2007). Lexical access, lexical representation, and vowel production. *Laboratory phonology*, 1–27.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of speech, language, and hearing research : JSLHR*, *47*(5), 1048–58.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology, 21*, 422–432.

Newman, R. S., Clouse, S. a., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142.

Norris, Dennis, James McQueen and Anne Cutler. (2003). Perceptual learning in speech. *Cognitive Psychology 47*, 204-238.

Nosofsky, Robert. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology, General 115*, 39-57.

University of California, Berkeley. Cognitive Science Program, & Ohala, J. J. (1984). *Explanation in phonology: opinions and examples*.

Ohala, John. J. (1989). Sound change is drawn from a pool of synchronic variation. In Leiv E. Breivik and Ernst H. Jahr (eds.), *Language Change, Contributions to the study of its causes*. [Series, Trends in Linguistics, Studies and Monographs No. 43]. Berlin: Mouton de Gruyter, 173-198.

Padgett, Jaye (2003a). Contrast and post-velar fronting in Russian. *NLLT* 21. 39–87.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119(4)*, 2382–2392.

Pardo, J. S. (2010). Expressing oneself in conversational interaction. To appear. In: E. Morsella (Ed.), *Expressing oneself/expressing one's self* (pp. 183–196). Taylor & Francis.

Peterson, Gordon E. & Harold L. Barney (1952). Control methods used in a study of the vowels. JASA 24. 175–184.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological studies in language*, *45*, 137-158.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory phonology*, *7*, 101-139.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. Psychological Science, 12, 348–351.

Sanders, N., Padgett, J.(2008b). Exploring the role of production in predicting vowel inventories. Paper presented at the 82nd meeting of the Linguistic Society of America.

Shankweiler, D., Strange, W., and Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In 'Perceiving, Acting and Knowing: Toward an Ecological Psychology', edited by R. Shaw and J. Bransford (Erlbaum, Hillsdale, NJ), pp. 315–345.

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–6.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–82.

Uchanski, R. M. (2005). "Clear speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishers, Malden, MA), p. 207–235.

Uther, M., Knoll, M. A., and Burnham, D. (2007). "Do you speak E-NG-L- I-SH? Similarities and differences in speech to foreigners and infants," *Speech Communication 49*, 1–7.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (1976). What information enables a listener to map a talkers vowel space? *J. Acoust. Soc. Am. 60*, 198–212.

Vroomen, J., & Baart, M. (2009). Recalibration of Phonetic Categories by Lipread Speech: Measuring Aftereffects After a 24-hour Delay. *Language and Speech*, *52*(2-3), 341–350.

Vroomen, Jean, Van Linden, S., De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–7.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, *23*(3), 247–274.

Wedel, AB. (2004). Self-organization and categorical behavior in phonology. Dissertation. UC Santa Cruz.

Wedel, A. (2004, July). Category competition drives contrast maintenance within an exemplar-based production/perception loop. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology* (pp. 1-10). Association for Computational Linguistics.

Wedel, Andrew, Kaplan, A., & Jackson, S. (2012). Meaning, Language Use and the Maintenance of Sound Category Contrast : A Corpus Study, (M.S.), 1–12.

Wedel, A. B., Kaplan, A., & Jackson, S. (n.d.). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *85721*(415), 1–20. Retrieved from http://dingo.sbs.arizona.edu/~wedel/publications/PDF/WedelKaplanJackson_2012_final.pdf

Wright, R. (2004). Factors of lexical competition in vowel articulation. In Local, J. J., Ogden, R., and Temple, R., editors, *Papers in Laboratory Phonology, volume VI*, pages 26–50. Cambridge University Press, Cambridge.

Zanone, Pier. G., and Kelso, J. A. Scott. (1992). The evolution of behavioral attractors with learning, nonequilibrium phase transitions. *Journal of Experimental Psychology, Human Perception and Performance 18*, 403-421.

Zanone, P. G., and Kelso, J. A. S. (1997). The coordination dynamics of learning and transfer, Collective and component levels. *Journal of Experimental Psychology, Human Perception and Performance 23*, 1454–1480.