THE GENERALIZABLE NATURE OF LEXICAL RETUNING

By

Scott Nelson

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Linguistics – Master of Arts

2019

ABSTRACT

THE GENERALIZABLE NATURE OF LEXICAL RETUNING

By

Scott Nelson

Auditory speech identification has been observed to be influenced by both lexical and visual information. Perceptual learning experiments have used two unique paradigms to test how each of these information sources affects the identification of ambiguous stimuli. In both cases, listeners are more likely to identify ambiguous stimuli in the direction of the disambiguating information they receive. It has been further argued that the resulting effects are the same and can be traced back to the same general speech perception mechanism. Despite this claim, there have been conflicting results in regards to generalization. Lexically induced perceptual learning has been observed to generalize to new contexts, while visually induced perceptual learning has been observed to be context dependent.

While the difference in these observed results could be explained by the information source (lexical vs. visual), there are also crucial differences in the experimental designs that may offer a better account. The training stimuli set for lexically induced perceptual learning experiments includes many unique tokens that are presented one time each. For visually induced perceptual learning experiments, the training set includes just one unique token presented multiple times. Listeners therefore only receive type variation in the lexically induced perceptual learning experiments. Crucially, type variation has been observed to be necessary for learning linguistic patterns and therefore may explain the differences in observed results between the two paradigms.

This current study uses three new experiments to study the generalizable nature of lexically induced perceptual learning. The results corroborate the idea that generalization of the effect to new contexts is possible in lexically induced perceptual learning experiments when listeners are trained with type variation, but when type variation is eliminated the ability to generalize the effect to new contexts is no longer observed.

ACKNOWLEDGEMENTS

I am heavily indebted to Dr. Karthik Durvasula and would first and foremost like to extend the most heartfelt thank you to him for all of the help and guidance he has provided me with during my time at Michigan State University. This thesis would not have been possible without his encouragement and support, and his technical and theoretical knowledge helped shape it into its current form. I am lucky to have had a mentor who I could discuss technical phonetic measurements with one day and then spend an hour discussing the philosophy of science the next. And I will always remember that "it's just simple physics."

I would also like to thank Dr. Yen-Hwei Lin who, in my first ever phonology class, said to me, "you think like a linguist." Those words have stuck with me since that moment and her critical advice and support have been crucial to my growth as a researcher Dr. Alan Beretta deserves a special thank you as well for his guidance and willingness to sit on my committee.

I would also like to acknowledge Dr. Suzanne Wagner, Dr. Alan Munn, and Dr. Marcin Morzycki for teaching me how to be a linguist both in and out of the classroom. My peers in the graduate program also deserve a special mention. Cara Feldscher, Chad Hall, Alex Mason, Monica Nesbit, and especially Kaylin Smith have made life more enjoyable even on the darkest days.

It would be foolish of me to sound off without thanking my parents, Bob and Kris Nelson. They have supported me unconditionally throughout the unconventional and every-changing path that has lead me here. Finally, I want to thank Kameron Chauvez, Shelbye Herbin, and Benjamin Linus Nelson for being the best friend, partner, and dog during this stressful (but rewarding!!) moment of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A theory of how listeners process ambiguous speech sounds is an important part of a larger theory of speech perception. If the ultimate goal of a listener is to reverse infer some discrete underlying representation from the continuous speech signal (Gaskell and Marslen-Wilson, 1996, 1998; Gow, 2003; Durvasula and Kahng, 2015, 2016; Durvasula et al., 2018), receiving ambiguous input increases the difficulty of this task. When the auditory input is sufficiently ambiguous, a listener may search outside the auditory domain in order to find cues that may help provide disambiguating information. Past research into these types of phenomena have shown that identification can be affected by lexical (Ganong, 1980) or visual (McGurk and MacDonald, 1976) information. More recent experimental paradigms have used either lexical or visual information to show that listeners' identification responses can be predictably biased towards a specific side of a perceptual boundary based on the disambiguating information they receive. Lexical retuning uses lexical information to bias listeners responses in a specific direction (Norris et al., 2003), while audio-visual recalibration relies on visual information to guide a listener towards identification (Bertelson et al., 2003). In both cases, it is non-auditory information that ultimately leads the listeners to their identification of the target sounds.

One question that falls from this fact is whether all forms of disambiguating information are treated equally by listeners. Is it the case that certain types of input are given a privileged status when decoding the speech signal? Despite relying on different information sources, lexical retuning and audio-visual recalibration often appear to have very similar consequences. In fact, Van Linden and Vroomen (2007) ran a series of five experiments to show that in all cases, both paradigms lead to results that are similar to one another. Due to these findings, they claimed that both identification strategies were part of the same underlying perception mechanism. On the surface, these results seem to suggest that all disambiguating information may be treated equally, but a more fine-grained analysis shows the necessity for further inquiry.

The validity of Van Linden and Vroomen's (2007) claim has since been questioned due to experimental results in both domains. A shortcoming of the original study was that, despite running five experiments, they were unable to test every dimension in which the two experimental paradigms had potential for variation. One conflicting dimension that their experiments did not cover is in the realm of generalization. Using lexical retuning experiments, it has been argued that the perceptual retuning effect is able to generalize over features (Kraljic and Samuel, 2006; Durvasula and Nelson, 2018), across syllabic position (Jesse and McQueen, 2011), and to previously unheard words (McQueen et al., 2006a). In contrast, audio-visual recalibration results suggest that the recalibration effect is strongly contextually bound (Reinisch et al., 2014). Reinisch et al. (2014) found that the phonological environment of the training and testing segments, the phonetic cues, and the segment itself needed to be identical in order for the effect to occur, ultimately demonstrating that generalization was not possible under their experimental conditions.

The results regarding generalization question the original claim put forth by Van Linden and Vroomen (2007). Perhaps lexical information is more privileged than visual information when tested over the proper dimension. Based on the data Van Linden and Vroomen (2007) had, their conclusion was reasonable, but new evidence may require an update to the claim that lexical retuning and audio-visual retuning are tapping into the same perceptual mechanism. Despite results over certain dimensions looking the same, it is still possible that the two are tapping into different levels of representation; which could therefore explain the difference in generalizability. A deeper inspection of lexical retuning could shed greater insight on what types of representations the effect is tapping into and whether or not these are more abstract (possibly, phonological) in nature.

The option to place a distinction between high-level lexical information and low-level visual information is enticing, but a counter argument to this source-based story can be made if we consider the stimuli used in these retuning/recalibration studies. In lexical retuning experiments, the training stimuli is a set of lexical items, each of which the listener hears once. In audio-visual recalibration experiments, there is typically only one training item that a listener is exposed to multiple times. The key here is that even when listeners are exposed to the same number of total stimuli, the number

of contexts in which they are exposed to the ambiguous input varies between the two paradigms. It is possible that the reason that learners in the audio-visual recalibration paradigm are unable to generalize the effect is because they internalize the ambiguity as an idiosyncratic pronunciation of that single string. Contextual variation has be shown to be important in generalizing linguistic patterns (Gerken and Bollt, 2008; Denby et al., 2018). It is therefore important to see what effect stimulus variation has on lexical retuning before exploring representations further. In order to do this, I will use a series of new experiments to further test the generalizable nature of the lexical retuning effect.

The remainder of this thesis will be formatted as follows. In chapter 2 I will give an explanation and background of lexical retuning and audio-visual recalibration. This will include detailed descriptions of the experimental paradigms as well as a comparison between the two. Part of this chapter will cover generalization as it relates to the overall perceptual learning literature. This will help to set up the motivations for the experiments described in chapter 3, chapter 4, and chapter 5. Chapter 6 will contain a general conclusion and recap what role each experiment will have on theories of speech perception and our overall understanding of perceptual retuning and recalibration.

CHAPTER 2

BACKGROUND

2.1 Lexical retuning

The lexical retuning paradigm was developed by Norris et al. (2003) in order to test the effect of lexical feedback on adapting to an unusual speaker. In this original set of experiments, they predicted that, because of the bias that gets induced *vis-à-vis* the "ganong effect" (Ganong, 1980), listeners' identification responses to an f~s continuum would vary depending on in which words a listener heard an ambiguous blend of [f] and [s] ([?$_{fs}$]).[1] For their experiment, they created a list of 40 Dutch words containing either /f/ or /s/ (20 of each) that were non-minimal pairs for the opposing sound (e.g., *witlof* "chicory"; *naaldbos* "pine forest") and a blended 41-step continuum of [f] and [s] sounds spliced onto an [ɛ] vowel. They chose 14 steps from the continuum to use in a pre-test. The results of the pre-test were used to find the maximally ambiguous point on the continuum and five steps to be used in the testing portion of the main experiment. The fricative portion of the maximally ambiguous point was then used as [?$_{fs}$] in the training portion of the experiment.

The overall design of the experiment was a lexical decision task (LDT) followed by phonetic categorization of the five steps identified in the pre-test. The LDT was broken into three groups. Each list had 100 real words and 100 nonce words. Of the 100 real words, 40 of them were the words that contained either /f/ or /s/. One group of participants heard [?$_{fs}$] in the /f/-words, a second group heard it in /s/-words, and a final group heard it in phonotactically licit Dutch nonce-words. The group who heard [?$_{fs}$] in nonce words also heard the regular /f/ and /s/ words with their standard unambiguous pronunciations.

It was predicted that participants who heard the ambiguous segment in words normally containing /f/ would give more "f" responses on the phonetic categorization task, especially for the

[1] Hereafter, I will refer to a sound ambiguous between two sounds [X] and [Y] as [?$_{XY}$].

points on the segment that were already ambiguous. Similarly, listeners who heard [?$_{fs}$] in words that normally contained /s/ would do the opposite and give more "s" responses. The third group who heard the ambiguous sound in nonce words were predicted to act as a control group. Their responses were therefore predicted to lie in the middle of the other two groups.

Results from their first experiment are shown below in Figure 2.1. "?f" represents the group that heard [?$_{fs}$] in /f/-words and "?s" represents the group that heard [?$_{fs}$] in /s/-words. The "non" refers to the control group that heard [?$_{fs}$] in nonce words. The x-axis is step number on the f~s continuum and the y-axis is how many "s" responses were given. The "?f" group gave significantly fewer "s" responses and then "?s" group gave significantly more "s" responses. There was no statistically significant difference between the control group and either of the other groups.

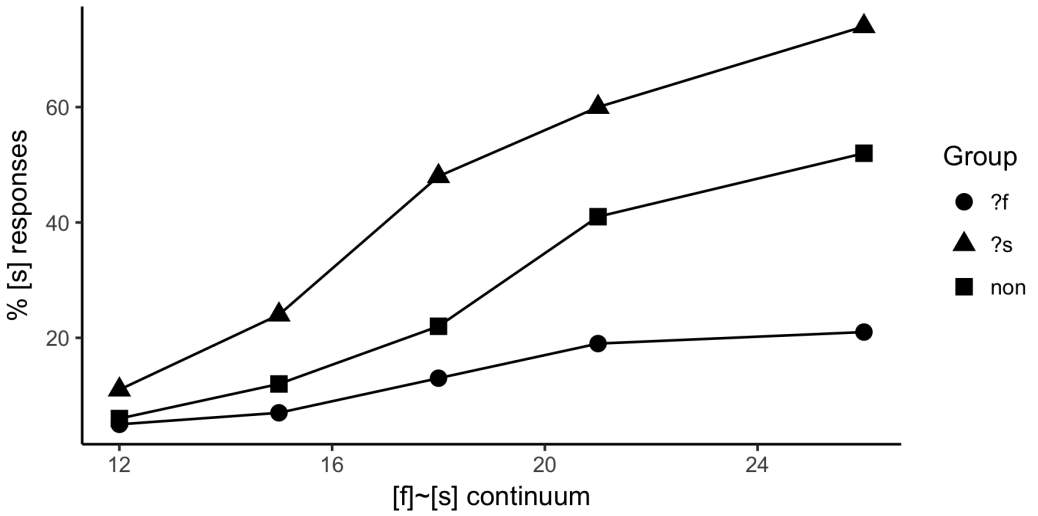

Figure 2.1: **Results adapted from Norris et al. (2003)**

These results suggest that lexical information can modulate the way in which ambiguous stimuli are categorized. Using a second experiment in their original study, Norris et al. (2003) confirm that their results cannot be explained by selective adaptation (Eimas and Corbit, 1973; Samuel, 1986). [2] With both of these results in hand, it is proposed that during the speech perception process listeners are getting lexical feedback not for on-line recognition but instead for learning (Norris et al., 2003,

[2]Selective adaptation refers to the phenomenon where repeated presentation of a stimulus results in temporary insensitivity to that stimulus.

p. 233-234).[3] Therefore, this paradigm and the resulting effect have been frequently referred to as "perceptual learning" (Norris et al., 2003; Eisner and McQueen, 2005, 2006; Kraljic and Samuel, 2005, 2006; Samuel and Kraljic, 2009). Since the original study, other segments beyond fricatives have also been shown to work within the paradigm. Stops (Kraljic and Samuel, 2006), liquids (Scharenborg et al., 2011), and lexical tones (Mitterer et al., 2011), and phonotactic information (Cutler et al., 2008) have all been shown to elicit and effect using lexical retuning experiments, suggesting that the effect is generally not something specific to certain acoustic cues, but rather an adaptation response to ambiguous, but categorizable, speech.

If lexical retuning is modeled as a learning effect then it is fairly clear how it may lead to higher level abstraction/generalization. The representation for the two segments in competition are constantly being updated, and the sampling distribution for an individual changes depending on which words they heard the ambiguous segment in. A listener learns in the sense that they need to learn the distribution for the speaker they are hearing (Kleinschmidt and Jaeger, 2015).

2.2 Audio-visual recalibration

During the same time frame that the lexical retuning paradigm was beginning, the audio-visual recalibration paradigm was developed by Bertelson et al. (2003). Here, it was knowledge of the McGurk effect (McGurk and MacDonald, 1976) that drove its creation. McGurk and MacDonald (1976) observed that when a listener gets visual (lipread) information that contradicts the auditory input, their identification of the auditory segment can be modulated by visual (lipread) information. [4] Using this general phenomenon, Bertelson et al. (2003) tested what would happen

[3] While the use of on-line here may seem non-standard, the basic idea is that they argue strongly against models of speech perception with lexical feedback in (Norris et al., 2000) and therefore need to make this differentiation between recognition and learning in order to maintain their strict feedforward model. They believe that the learning happens over time and therefore is qualitatively different than on-line feedback for recognition.

[4] The effect is more prevalent when using a non-labial segment for the visual presentation accompanied by a labial auditory segment than the opposite scenario. This is briefly noted by Bertelson et al. (2003), but because they were interested in the "aftereffects" (learning/recalibration rather than initial response), they did not find it worrisome for their experiment.

when participants are presented ambiguous audio alongside unambiguous visual input using a nine-step b~d continuum and audio/video recordings of a male native Dutch speaker pronouncing /aba/ and /ada/.

Each participant took a pre-test prior to the main portion of the experiment to find their individual maximally ambiguous point ([$?_{bd}$]) along the continuum. The audio token corresponding to [$?_{bd}$] and the two steps ±1 step from the token were used in the main experiment. [$?_{bd}$] paired with unambiguous visual cues were used as training items and all three steps were used in a post-training phonetic categorization task (two alternative forced choice; "aba" or "ada"). The main experiment consisted of 16 blocks. Each block consisted of eight exposure trials where participants were presented with one of the two visual cues paired with [$?_{bd}$], followed by six phonetic categorization tasks (two each of the three steps determined by the pre-experiment). Eight of the blocks used the /aba/ visual cue and the remaining eight used the /ada/ visual cue. Listeners were therefore exposed to both training segments.

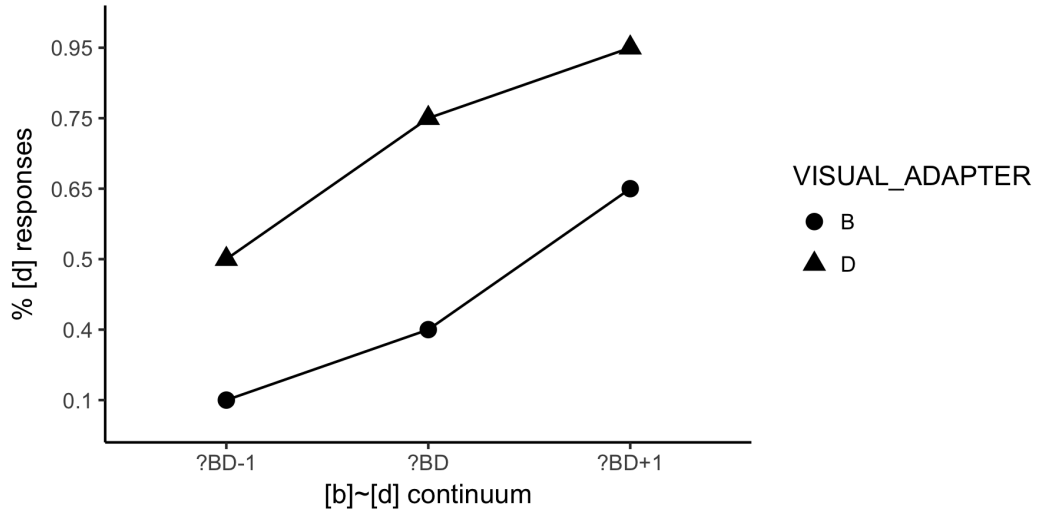

Figure 2.2: **Results adapted from Bertelson et al. (2003)**

Results from Bertelson et al.'s (2003) initial study on audio-visual recalibration are presented in Figure 2.2. The "B" visual adaptor group corresponds to participants who saw unambiguous visual input of a /b/ simultaneously with the ambiguous [$?_{bd}$] audio segment. The "D" visual adaptor

group saw /d/ visual input, along with the ambiguous sound. Because the phonetic categorization task consisted of the most ambiguous sound on a /b/~/d/ continuum for each individual participant and the steps ± 1 from that step on the continuum, the graph below is normalized across speakers and does not constitute any one specific area of the continuum. Speakers who saw the /d/ visual input were more likely to respond "d" afterwards and speakers who saw /b/ visual input were more likely to respond with "b".

The results presented demonstrate that the identification of ambiguous auditory input can be biased by visual information. Bertelson et al. (2003) interpret these results fairly generally and conclude that when cross modal biases occur, recalibration of some sort must occur.

2.3 Comparison of the two paradigms

While the two paradigms use different methods to disambiguate the ambiguity of auditory stimuli, they appear at a cursory glance to be inducing a similar perceptual shift. In both cases, when presented with an unambiguous anchor (either non-minimal pair lexical item or clear visual presentation), listeners adapt their responses in the direction of the unambiguous anchor. Despite this similarity, there are a number of dimensions in which the two strategies vary and lead to skepticism about whether or not these are parts of the same general speech perception mechanism. These differences are outlined in Table 2.1 and discussed further below.

	Lexical Retuning	Audio-Visual Recalibration
Processing Strategy	Top-down	Bottom-up
Speech Type(s)	Ambiguous	Clear or Ambiguous
Stability	Relatively long lasting	Relatively short lasting
Presence of opposing phoneme	Yes	No
Experiment Design	Between Subjects	Within Subjects

Table 2.1: **Comparison of lexical retuning and audio-visual recalibration paradigms**

The first way to compare the two is with their processing strategies. While it can be argued that the use of visual information/lip-reading is a bottom-up strategy, it's clear that the use of lexical information relies on top-down processing. This difference could be what leads to variation between the two paradigms in the ability of their effects to generalize. There is also a strong difference

in what types of speech can elicit each effect. Lexical retuning can only work on ambiguous speech. If clear speech is used then there is no need for any disambiguation strategy. On the other hand, McGurk effects can alter the perception of clear speech (McGurk and MacDonald, 1976). This suggests that visual information is potentially tapping into a different underlying mechanism, potentially one that is autonomous. Further support for the autonomous view comes from Baart and Vroomen (2010) who observed that the effects of audiovisual recalibration persisted even when participants were simultaneously given tasks that put a strain on working memory. These findings put audiovisual recalibration in a similar category as selective adaptation which has also shown to be an automatic process that is unaffected by memory load (Samuel and Kat, 1998). Because Norris et al. (2003) ran a second experiment to show that lexical retuning effects could not be explained by selective adaptation, this gives additional substance to the hypothesis that audiovisual recalibration and lexical retuning are not tapping into the same general perceptual mechanism. If the two effects target unique spots along the perceptual stream then this would once again give credence to why one allows for generalization and the other does not.

Another area in which empirical results for each paradigm have conflicted is stability or how long each effect lasts. Vroomen et al. (2004) observed that the recalibration effect induced by audio-visual information was relatively short lasting. After hearing 50 exposure stimuli, the effect wore off after only 6 post-training tests. In the realm of lexical retuning, independent results suggest that the effect was relatively long lasting. Kraljic and Samuel (2005) observed that it could last up to 25 minutes while Eisner and McQueen (2006) found the effect to last for 12 hours regardless if the participant slept or not. One could argue for an account of these results based on where along the perceptual stream each effect is being integrated. The long-lasting nature of the lexical retuning effect could be due to it affecting a higher-level representation, while the audio-visual information may simply bias the recognition process. In other words, the visual information changes the way a sound is perceived at the input, but the lexical information cannot alter the perception until later on down stream after the high level information has been obtained.

Additionally, there is the learning aspect to consider. Lexical retuning experiments were

originally developed to see how speakers adapted to novel accents. Norris et al. (2003) argued that higher level information could be used to provide feedback if learning was the ultimate goal.[5] Someone pronouncing a word strangely may necessitate learning of the speaker's idiosyncratic mappings from surface to underlying forms, but conflicting audio and visual information does not require this since it does not require the same type of mapping. Speculating on this may be unnecessary, though, as there are other confounds between the previously described experiments (such as silence following the training) that may also explain the difference in results.

There are a few general design differences between lexical retuning and audio-visual recalibration experiments as well. The presence or absence of the opposing phoneme in training conditions is one of them. Lexical retuning experiments include both segments from the continuum while audio-visual recalibration experiments have only the phoneme being tested presented in each specific training-testing block. Additionally, because the design of the audio-visual recalibration experiments is blocked in a way that participants are exposed to both phonemes as the ambiguous segment, this means that they are constantly reinterpreting the same ambiguous sound as both discrete categories in a way that participants of lexical retuning experiments are not.

There also has been variation in the segment type and syllabic position used throughout the two paradigms with only brief mentions throughout the literature suggesting that this could be causing the difference in results. Audio-visual recalibration experiments present the same training token throughout the entire block. Since they are only hearing the ambiguous auditory segment in one environment, it is possible that participants may just consider the ambiguity to be unique to that string. Due to the design of lexical retuning experiments, participants hear the ambiguous segment in multiple words. The stimulus variation in lexical retuning may lead listeners to the conclusion that this is a general principle of the speaker and not just an idiosyncratic pronunciation of one string. If this is the case, then it may explain why lexical retuning experiments have shown

[5] As discussed previously in footnote 3, this was primarily due to their Merge model of speech perception which was strictly feedforward (Norris et al., 2000) The learning aspect was added to differentiate between different types of on-line feedback. They ultimately argue that on-line feedback has no benefit to perception and is only used when learning is involved.

generalization effects in ways that audio-visual recalibration have not.

To test how lexical retuning and audio-visual recalibration performed head to head, Van Linden and Vroomen (2007) created a series of five experiments where the same ambiguous segment [?$_{tp}$] was used in both tasks. In both cases, the ambiguous segment appeared in coda position. The five experiments that were run can be broken down as follows: Experiment 1 resulted in both paradigms having similar size effects; Experiment 2 found that the effect dissipated relatively quickly for both paradigms; Experiment 3 observed that the presence of the opposing phoneme increased the size of the effect for both paradigms; Experiment 4 found that a three minute silence period between training and testing did nothing to affect the size/stability of the effect; Experiment 5 resulted in listeners who were trained on both of the phoneme categories not performing any differently than those who were trained on only one category.[6]

Their takeaway from all experiments is that lexical retuning and audiovisual recalibration show similar aftereffects and therefore possibly represent the same general mechanism. This was taken as the standard view and even echoed in Samuel and Kraljic's (2009) general overview of the phenomenon. The problem with assuming that the results from Van Linden and Vroomen's (2007) study indicate an overall similarity between the two perceptual effects is that they only tested certain dimensions. As mentioned previously, there have been additional conflicting results in regards to generalization within both paradigms. These must be addressed so that we can better evaluate the overall nature of perceptual recalibration.

2.3.1 Generalization

Generalization in lexical retuning experiments has been shown in a few different ways. Going back to the original experiment (Norris et al., 2003), one can make the argument that lexical retuning has always had the effect of promoting generalization. In this experiment, the training word lists

[6]Some of these results are in conflict with prior results in the literature. Some possibilities for this conflict may be syllabic position or consonant type. Due to the different labs working within each paradigm, there was variation in the types of segments and syllabic position used for prior comparisons. These new results from Van Linden and Vroomen (2007) may be observed due to the better control of these two variables.

that were used contained a multitude of different vowel-fricative bigrams while the testing string was always from an [ɛf]~[ɛs] continuum. Only one of the training words contained an [ɛ] vowel. Therefore, literally only one instance of the training stimuli matching the test stimuli is enough to induce the lexical retuning perceptual shift, or listeners were able to learn a generalized pattern that was independent of the environment of the ambiguous segment[7].

Additionally, Jesse and McQueen (2011) observed that when listeners were trained with the ambiguous segment in coda position, they transferred the effect when tested on segments from a [fø]~[sø] continuum. These results suggest a generalization to a position-independent representation. Finally, generalization to a featural level has been shown for both stops (Kraljic and Samuel, 2006) and fricatives (Durvasula and Nelson, 2018). In both instances, training and testing environments were different, suggesting that the retuning effect was operating over a more abstract representation than just auditory features.

In the realm of audio-visual recalibration the opposite has been found. Reinisch et al. (2014) observed no evidence of generalization in a series of three experiments that suggested that both training and testing environments needed to be identical in order to induce the perceptual recalibration. They interpret their results to mean that the recalibration is specific to phonemes, acoustic cues, and the phonological context, but state that, "This conclusion rests, however, on the assumption that the visually-guided recalibration reflects a general speech-perception mechanism (an assumption empirically supported by Van Linden and Vroomen (2007)" (p. 104). The general speech perception mechanism alluded to here is one in which audio-visual retuning is the same as lexical retuning. A driving factor for this thesis is to see whether or not this assumption should continue to be made. I will use the lexical retuning paradigm to pursue this further.

There are two issues that I am specifically going to look at relating to this assumption. Van Lin-

[7] A potentially more interesting question is whether or not listeners would be able to generalize the pattern to an environment they didn't hear despite hearing multiple environments in the training stage. In other words, if they hear the ambiguous segment in multiple environments but not every environment they're aware of, would they generalize to the non-trained environments? Depending on how listeners pattern, this could be a way to further pursue questions of representation and possibly even learnability

den and Vroomen (2007) did not include any experiments that would test whether or not the two paradigms resulted in generalization in their comparison study. To see if indeed the two paradigms use the same general speech perception mechanism, one needs to establish that both of them result in a similar degree of generalization under similar conditions. One way to make this comparison is to use an experimental setup similar to Reinisch et al. (2014), but within the lexical retuning paradigm. In their experiment, they strictly controlled the phonetic environments of the training and testing conditions. No such study using the lexical retuning paradigm has explicitly done the same thing. Furthermore, one possible reason that lexical retuning experiments have resulted in more generalizability than audio-visual recalibration experiments is the increased amount of stimulus variation present. The lack of such stimulus variation in audio-visual recalibration may lead participants to view the ambiguous sound as idiosyncratic to the particular string they hear it in. Probing if generalization is present in lexical retuning experiments in the absence of stimulus variability would allow us to understand if such differences in stimulus variability could be the source of the observed discrepancy between the two paradigms.

If we look outside of the perceptual learning paradigm, it has been shown that linguistic generalizations require type experience and not token experience in order to be learned. Gerken and Bollt (2008) present results that suggest that 9-month old infants are able to generalize a constraint that says heavy syllables should be stressed to novel stimuli when presented with three unique training items, but failed to do so when presented with one unique training item presented multiple times. Denby et al. (2018) draw a similar conclusion using artificial language learning experiments. They looked at listeners abilities to learn gradient phonotactic patterns and found that contextual variability (type frequency) significantly affected learning while the number of times an exemplar was repeated (token frequency) had no such effect. Both of these results taken together suggest that in order to generalize linguistic patterns, type experience is needed.

Three experiments relating to these issues will be used to 1) more carefully test the generalizable nature of lexical retuning and 2) see what effect the type vs. token experience has on it. Experiment 1 is primarily a replication of Norris et al. (2003) with a slight modification of the control group. This

is used to corroborate the idea that lexical retuning can be learned from many conditions in training to one condition in testing. Experiment 2 is an attempt to replicate some of the training/testing environments used in Reinisch et al. (2014), but within the lexical retuning paradigm. In this experiment, the training condition is limited to one vowel and is unique from the vowel used in the testing condition. This is used to see if lexical retuning shows generalization effects when strictly controlling the training and testing environment. In this sense it may give a stronger indication at the overall generalizable nature of the lexical retuning effect. Experiment 3 uses a similarly restricted training/testing environment as Experiment 2, but reduces the training stimuli set in order to see how the absence of type variation affects generalization. This, even more closely than Experiment 2, replicates the training and testing that participants underwent in Reinisch et al.'s (2014) experiments. The results from all three experiments will give us a better idea about how and why the generalization facts vary between lexical retuning and audio-visual recalibration.

CHAPTER 3

EXPERIMENT 1: REPRODUCING GENERALIZATION IN LEXICAL RETUNING EXPERIMENTS

In Experiment 1 the goal is to confirm that the baseline generalization that is seen in Norris et al.'s (2003) experiment can be replicated. This experiment looks at training conditions with multiple unique vowels as well as variation in syllabic position. The testing block of the experiment has the target fricative in onset position. Since it was shown that lexical retuning can transfer its effect from coda to onset but not vice versa (Jesse and McQueen, 2011), this allows for the use of both syllabic positions and therefore increases the set of potential words to use in the training stimuli. One significant difference between Experiment 1 and the Norris et al. (2003) experiment is in the control group. In their experiment, Norris et al. (2003) include nonce words containing an ambiguous segment and normal words containing [f] and [s] in their control group's LDT wordlist and argue that because they do not form real words then the retuning effect will not occur since there is no lexical information to tap into. Since this was never clearly confirmed, there is a possibility that the presence of the ambiguous segment still had an influence on the control group despite appearing in nonce words. To err on the side of caution, the control group for Experiment 1 contains no instances of an ambiguous segment, but rather just the clear [f] and [s] tokens and the same filler words as the test group. This will hopefully control for any subtle effect of the ambiguous segment for those in the control group.

The other primary departure from previous work using lexical retuning is the addition of a pre-LDT phonetic categorization task while also keeping the post-LDT phonetic categorization task. This allows for before-after comparisons to be made within both the control and test groups, as well as the traditional 'After"-"After" comparison that is typically used in lexical retuning analysis. I also add a second between-group comparison which is the difference from before to after for both groups and argue that this is a more insightful comparison.

3.1 Method

3.1.1 Participants

Seventy-three undergraduate students from Michigan State University (Mean Age = 21; 57 female, 4 unreported gender) received either course credit or a small monetary reimbursement ($10) for participating in the study. All speakers identified as American English speakers and did not report any hearing problems.

3.1.2 Design

The experiment consists of three blocks. First, participants will be given a phonetic categorization task. They will hear various steps from a blended continuum of [f] and [s] segments that precedes an [i] vowel. This will be used as a baseline measurement for analysis. A training phase follows in the form of an auditory lexical decision task (LDT) which is used to introduce participants to the ambiguous segment that will trigger the perceptual learning. The LDT is then followed by a second phonetic categorization task, identical in form to the baseline test. The flow of the experiment is visualized below in Figure 3.1.

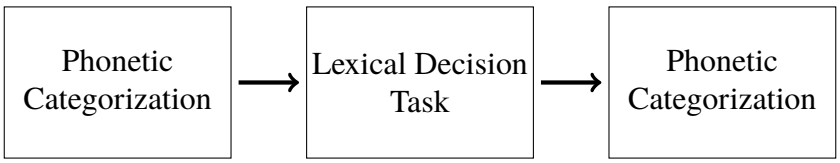

Figure 3.1: **General Experiment Design**

By adding the "Before" phonetic categorization block, it allows for both within- and between-subjects analysis. Traditionally, lexical retuning experiments have looked at between-subjects comparisons of post-LDT phonetic categorization tasks with no reference to a starting baseline (Norris et al., 2003; Eisner and McQueen, 2005, 2006; Kraljic and Samuel, 2005, 2006; McQueen et al., 2006b; Cutler et al., 2008; Sjerps and McQueen, 2010; Jesse and McQueen, 2011; Scharenborg et al., 2011; Reinisch and Holt, 2014; Mitterer et al., 2011, 2013, 2016). In these experiments,

it was tacitly assumed that each group was starting from the same baseline. Including both a "Before" and "After" categorization removes this assumption and better controls for any sampling idiosyncrasies. Each participant is assigned to one of two groups: $FISI_{test}$, $FISI_{control}$, which will be discussed in more detail below.

3.1.3 Materials

The lexical decision task uses a 150 word list containing 75 English words and 75 phonotactically licit English nonce words. Thirty-four of the English words are training items, while the remaining 41 and all 75 nonce words are used as filler. All 34 training tokens contain either an [f] or an [s] (each segment distributed equally) and do not form a minimal pair when replaced with the opposing segment (e.g., *beef, sing*). All the training words used were monosyllabic and contained one of nine different vowels. Nine of the 17 words for each segment had the crucial segment in the onset (word initial) position of the word. Test words were controlled for frequency using the Subtlexus corpus (Brysbaert and New, 2009). While the [f]-words were less frequent (12.85/million) than the [s]-words (20.77/million), no statistically significant difference between the log frequencies was found [t(28.3) = 0.48, p = 0.64]. The remaining 114 filler words contain no instances of [f s v z]. The real word fillers range in syllable count from 1-3 while the nonce word fillers range from 1-4.

The 150 words for the LDT were spoken by a female native American English speaker from Michigan. Each word was read aloud into a Logitech 980186-0403 microphone (frequency repsonse 100Hz–16kHz; -67dBV/ubar, -47dBV/Pascal ±-4dB) in a quiet room and recorded directly into Praat (Boersma and Weenink, 2016) at a sampling rate of 44.1 KHz. The speaker also recorded tokens of [fi] and [si]. These were used to make the continuum: fi~si. The creation process was done as follows: first, the tokens of [fi] and [si] were manually annotated in Praat (Boersma and Weenink, 2016) to mark the fricative and vowel portions of the token. From here, the entire process was automated using scripting. To make each continuum, equal amounts of the fricative portion of each token was spliced out (165 ms; normalized to 50dB). The amplitudes of the f/s pairs were then blended in 41 equal steps (e.g., step 1 was 100% [f] and 0% [s]; step 2 was 97.5% [f] and

2.5% [s]; . . . step 41 was 0% [f] and 100% [s]). The continuum was then re-spliced back onto the vowel portion from the [fi]. It is well known that that formant transition information in the vowel acoustics is a cue for place of articulation (Delattre et al., 1955, 1962), therefore this method may introduce a slight bias towards "f" responses, but in previous studies it has been shown to elicit normal response functions (Norris et al., 2003; McQueen et al., 2006b; Durvasula and Nelson, 2018)

The continuum was then used in a pre-test to find the most ambiguous step. Of the 41 total steps, 14 were chosen to use for phonetic categorization. These included steps 1 & 41 which were the 100% [f] and 100% [s] tokens, respectively, as well as every other step between steps 7-29. This meant that the majority of the tokens that participants categorized were from the ambiguous portion of each continuum. Thirteen American English speakers (Mean Age = 20.9 years; 8 female, 1 unreported gender) from Michigan State University separate from those in the main experiment participated in the pre-test for partial class credit.

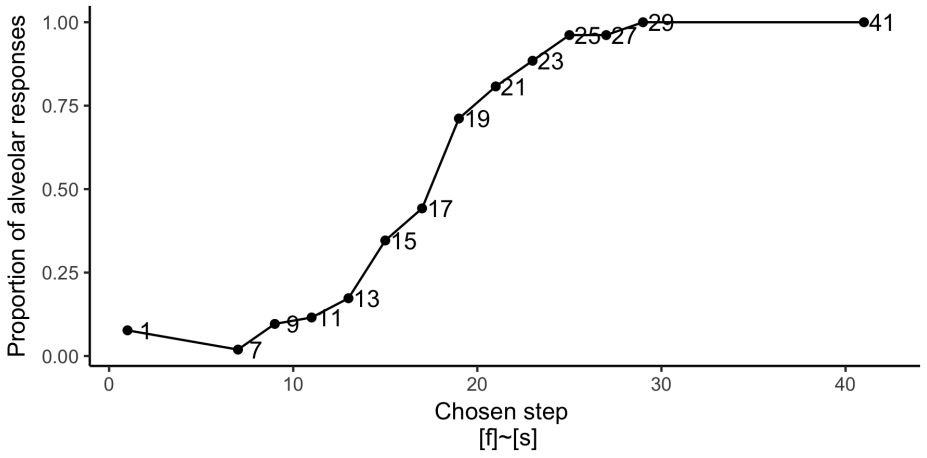

Figure 3.2: **Categorization results for the fi~si continuum pre-test**

Participants were tested using PsychoPy (Peirce et al., 2019). Each participant heard the 14 steps from the continuum four times each in random running order. A two alternative forced choice paradigm was used. After hearing a sound, the participant was instructed to use a computer mouse to indicate whether the sound they had just heard contained either "f" or "s". The mean response

for each step was calculated to create an identification response function. This resulted in the sigmoidal function expected for consonant segments (Liberman et al., 1957). The point closest to which the response function intersected with the 50% response rate was then interpreted as the most ambiguous point on the continuum. For the fi~si continuum, the 50% response rate lay in between steps 17 and 19 and therefore the fricative portion of step 18 was used to create [$?_{fs-i}$]. This was then used as a replacement for all of the [f] sounds in the LDT for the $FISI_{test}$ training group.

Praat (Boersma and Weenink, 2016) was once again used to manipulate audio stimuli, this time to create an altered versions of the LDT wordlist. The $FISI_{test}$ version of the list had all of the [f] tokens in the LDT replaced with [$?_{fs-i}$]. The $FISI_{control}$ wordlist was unaltered and therefore had no instances of an ambiguous token. To make the altered wordlist, the fricative portion of each word containing [f] was manually annotated and marked at points of zero crossing. Scripting was then used to automatically remove the original frication portion of the word and replace it with [$?_{fs-i}$]. The returned wordlist was then manually checked for naturalness based on author-intuition.

3.1.4 Procedure

The main experiment was also performed using PsychoPy (Peirce et al., 2019). Up to 12 participants were tested in the lab simultaneously. Prior to the experiment, participants were verbally instructed that they will be doing three tasks on the computer using various response mechanisms - the phonetic categorization (first and third tasks) requires the clicking of a mouse while the lexical decision task (second task) will require them to use the keyboard to give a yes/no response. Participants were also verbally instructed to answer as quickly and accurately as possible and to remain seated and quiet until everyone in the room had completed all three tasks. Participants wore over the ear headphones (Plantronics .Audio 355; 20Hz-20kHz response) throughout the entire experiment and specific instructions were visually presented to them on the screen before and during each task.

The phonetic categorization tasks were the same format as the pre-test described above: each iteration contained the same 14 set of steps played four time each in random running order.

Participants were randomly placed into one of the two groups. Both the $FISI_{test}$ and $FISI_{control}$ groups were played the same fi~si continuum. As before, participants were given a two alternative forced choice task (“f” or “s”) and instructed to use a computer mouse to click which sound they hear.

Figure 3.3: **Phonetic categorization Psychopy screen and instructions**

Upon completion of the first phonetic categorization task, participants underwent the lexical decision task. They were instructed through PsychoPy (Peirce et al., 2019) that they will now be hearing a series of words, one at a time, and have to decided whether or not the word they heard is a real English word. Additionally, they were instructed now to use the computer keyboard’s ‘a’ and ‘l’ keys to answer “no” or “yes” to the question, “Is this an English word?” An ‘a’ response corresponded to “no” and an ‘l’ response corresponded to “yes”. This information was constantly on screen as reference for participants.

Figure 3.4: **LDT Psychopy screen and instructions**

If a participant did not respond within 3.5 seconds of the onset of the sound, a new sound

was played and no response was recorded. Each word was presented in random running order. Depending on the group that an individual participant is assigned to, they were played the corresponding LDT list as described above. After completion of the LDT, participants were given the same phonetic categorization task that they did prior to it. The only difference from the previous phonetic categorization is that a new random running order of the stimuli was used.

3.2 Results

In the original lexical retuning experiments, Norris et al. (2003) set a criteria of 50% accuracy rate of the ambiguous segment in the LDT in order to keep a participants' data for analysis. This was to ensure that participants were recognizing it as a quality exemplar of whatever segment it was replacing. In this experiment, the criteria was expanded. Participants were required to score 50% or higher in accuracy for both of the training segments ([s] and either [f] or [$?_{fs}$]), as well as have an overall score of 50% or higher. Two participants ended up being removed from analysis. Overall, both groups had accuracy rates of 90%.

The within-subjects analysis was performed first. A one-tailed, paired t-test showed that there was a statistically significant decrease in alveolar responses from the "Before" phonetic categorization task to the "After" phonetic categorization task for the $FISI_{test}$ group [$\text{MeanDiff}_{\text{Before-After}}$ = -0.0873, t(35)=-4.8, p<.001]. This suggests that participants in the $FISI_{test}$ group altered their categorization of the fi~si after hearing [$?_{fs}$] in words normally containing [f] during the LDT. A one-tailed, paired t-test showed that there was also a statistically significant decrease in alveolar responses from the "Before" phonetic categorization task to the "After" phonetic categorization task for the $FISI_{control}$ group [$\text{MeanDiff}_{\text{Before-After}}$ = -0.0393, t(34)=-2.46, p<.01]. This is unexpected as there was no alteration to this group's LDT. One explanation for this result is the fact that the vowel portion of all of the phonetic categorization tokens was taken from a token of [fi]. It was predicted beforehand that if there was any bias, that it would be in the direction of "f" responses. These results seem to bear out that sentiment. The categorization functions for both of these comparisons can be seen in the upper left and upper right graphs of Figure 3.5.

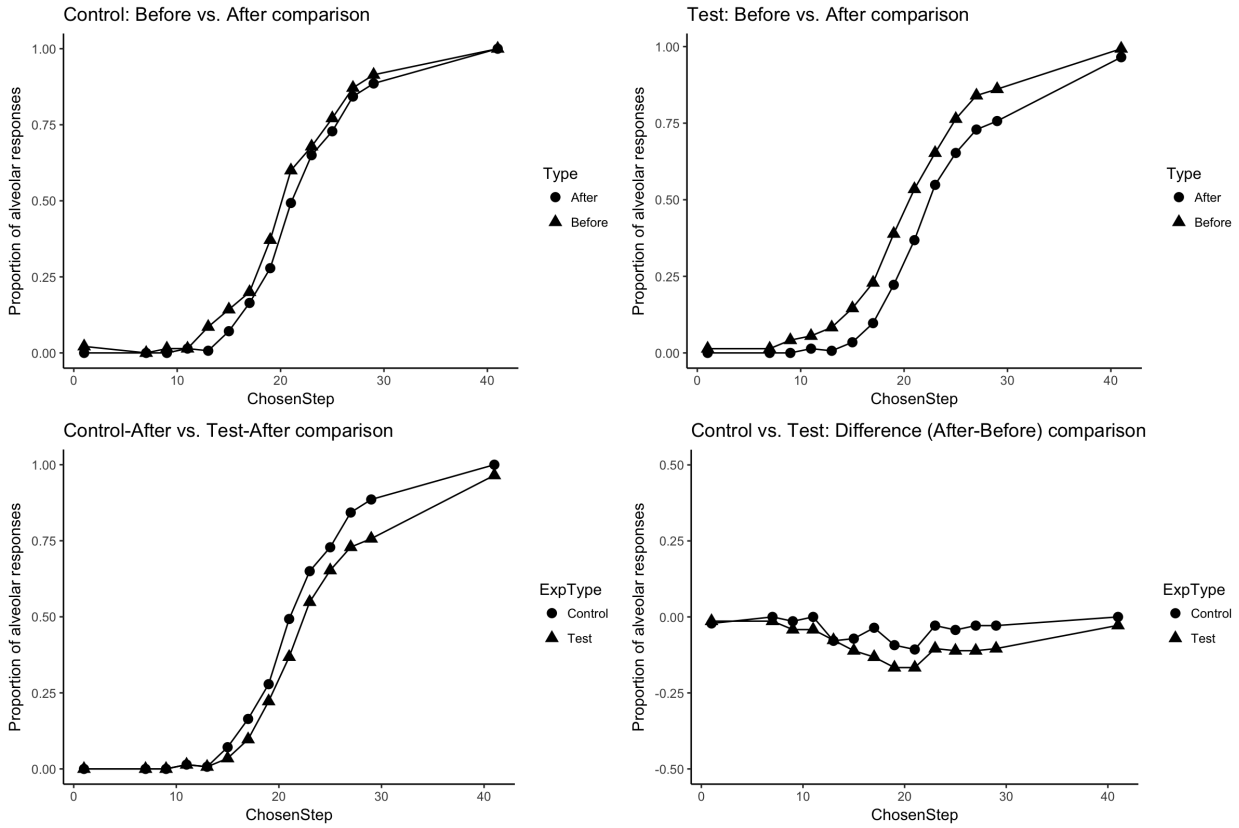

Figure 3.5: **Categorization results for Experiment 1**. Upper left is $FISI_{control}$ before vs after comparison; Upper right is $FISI_{test}$ before vs after comparison; Bottom left is the after responses of $FISI_{control}$ vs the after responses $FISI_{test}$; Bottom right is the difference between after and before responses between the two groups.

Since both of the within-subjects analyses proved to be significant, between-subjects tests were run. The first test was between the responses of the "After" phonetic categorization task for both groups. This is traditionally the comparison that is made in lexical retuning experiments. A one-tailed Welch test showed that there was a statistically significant difference in alveolar responses between the $FISI_{test}$ and $FISI_{control}$ groups [$\text{MeanDiff}_{\text{Test-Control}} = 0.053$, $t(67.9)=1.77$, $p=0.04$].[1] This result is slightly surprising as previously it has been observed that there was no difference between control and test groups (Norris et al., 2003). The design of this experiment also allows for a second between-subjects analysis. While the "After"-"After" comparison gives insight into

[1]Welch tests were used for the between subjects analysis since they do not assume homogeneity of variance between samples.

what participants are doing based on the experimental group they were assigned to, it does not take into account that different samples of the population may have different starting points for the phonetic categorization. The difference between the "After" results and the "Before" results can be taken, for both the test group and the control group, and used as comparison between the two groups. A one-tailed Welch test showed that there was also a statistically significant difference in alveolar responses between the two groups [$\text{MeanDiff}_{\text{Test-Control}} = 0.048$, $t(68.1)=1.99$, $p=0.026$]. The functions for both of these comparisons are in the bottom left and bottom right graphs of Figure 3.5.

3.3 Discussion

The results from Experiment 1 show that listeners give an increased number of "f" responses to a fi~si continuum after hearing [$?_{fs-i}$] in place of [f] tokens in an LDT. Since the [f] tokens were in both coda and onset position, and in a multitude of different vowel environments, it is possible that a generalization of the perceptual learning has occurred. But a statistically significant shift in the same direction was shown for the $FISI_{control}$ group as well. This indicates that there is some confound causing a shift from the "Before" phonetic categorization to the "After" phonetic categorization. It is possible that this is due to the formant transition information present in the fi~si continuum. For this reason, a comparison between the groups was necessary.

The "After"-"After" comparison that has typically been made when analyzing these types of experiments also showed a statistically significant difference in the expected direction (fewer "f" responses for the $FISI_{test}$ group). When looking at the results in the lower left graph of Figure 3.5, it can be seen that much of the difference between the two groups appears to be in the last three steps of the continuum. We expect the variation to be more present in the middle of the continuum. An inspection of the lower right graph of Figure 3.5 shows this. This graph shows the difference in responses within each group from the "Before" to the "After" phonetic categorization tasks. There is steady separation between the two groups starting at step 11 and going all the way through the rest of the continuum. Step 13 is the point in all the other graphs where the response rates

are in between 0-1 and therefore where we would predict less stable responses (i.e., this is where we should see the effects of the lexical retuning). The statistics confirm that there is a significant difference in the expected direction when comparing the two groups this way.

The overall takeaway from Experiment 1 is that there is some evidence for generalization within lexical retuning experiments. Participants were trained with the ambiguous token in a multitude of different vowel environments, but tested in only one. That being said, the training set of words for the LDT did include two words where [?$_{fs}$] was in coda position after [ɪ] and one time where it followed [i]. It also included two words where [?$_{fs}$] was in onset position before [ɪ]. So 5/17 words in the training set contained the same vowel (or one that is very similar) as in phonetic categorization testing. Therefore it could be argued that the perceptual learning followed from this smaller subset. To make the claim that it is in fact generalization to new contexts, stricter training and testing conditions are required. In Experiment 2, I will further probe the generalizable nature of lexical retuning by seeing whether the learning effect can generalize to a context that is not present in the training block.

CHAPTER 4

EXPERIMENT 2: GENERALIZATION WITH NON-IDENTICAL TRAINING AND TESTING CONDITIONS

Experiment 1 replicated previous results, and showed that the lexical retuning effect could be learned from many training environments and be used in one testing environment. It is hypothesized that this is because listeners learn a pattern and generalize it; however, there was some overlap in training/testing vowel environments, which places the generalization claim under dispute. It is possible that the subset of training words where the vowel context matched the testing condition may have been enough to induce the shift in categorization seen in Experiment 1.

Recall that context dependency was more strictly tested using the audio-visual recalibration paradigm. There it was shown that the perceptual learning effect was context dependent. One of the findings from Reinisch et al.'s (2014) study is that when participants are trained in the environment of an [i] vowel, they are unable to generalize the effect when tested on an [a] vowel. (i.e., /aba/ or /ada/ for training did not induce the perceptual recalibration shift for an [ibi]~[idi] continuum). There has been no study using the lexical retuning paradigm that has imposed as strict training and testing environmental restriction. For this reason, Experiment 2 uses the lexical retuning paradigm to see whether participants trained with an ambiguous segment in the environment of an [i] or [ɪ] vowel will show lexical retuning effects when tested on a [fa]~[sa] continuum. Since there is no overlap in vowel context between the training and testing conditions, a shift in the categorization by the test group would more explicitly support generalization of the effect over a context independent representation.

4.1 Method

4.1.1 Participants

Seventy-eight undergraduate students from Michigan State University (Mean Age = 19.5; 47 female,1 unreported gender) received either course credit or a small monetary reimbursement ($10)

for participating in the study. All speakers identified as American English speakers and did not report any hearing problems.

4.1.2 Design

The overall design for Experiment 2 was identical to that in Experiment 1. Experiment 2 has the same three part set up of pre-training phonetic categorization, training *vis-à-vis* an auditory lexical decision task, and post-training phonetic categorization. There are some more, however: the phonetic categorization continuum is now a fa~sa continuum and the training words in the lexical decision task are now limited to words that contain an [i] or [ɪ] vowel next to the target fricatives. These will be more explicitly outlined below. For this experiment, participants were assigned to one of two groups: $FASA_{test}$ or $FASA_{control}$.

4.1.3 Materials

An LDT was used with 150 words split into 75 English words and 75 phonotactically licit English nonce words. The 116 filler words (41 real/75 nonce) from Experiment 1 were once again used in the Experiment 2 list. The 34 training items are new. Since the goal of this experiment is to observe how the lexical retuning effect behaves when the phonological environment for the training and testing vary, all the training items have the crucial segments that are situated next to an [i] or [ɪ] vowel. Expanding the criteria beyond just the [i] vowel was necessary in order to obtain a large enough training sample. The choice of [ɪ] was due to it being the most phonetically similar segment in American English (Hillenbrand et al., 1995). Phonologically, the [i]–[ɪ] distinction is typically embodied over a contrast of ATR or length. In this study, the goal is to use phonological distance as a comparison and therefore only the height and front-back features are of interest. Therefore, we can use the [ɑ] vowel for contrast since it is maximally different from [i ɪ] on both of those dimensions.

All 34 training tokens once again contain either an [f] or an [s] (each segment distributed equally) and do not form a minimal pair when replaced with the opposing segment. For both segments,

13 of the tokens have the crucial segment in onset position. It appears in coda position for the remaining four. Each segment has 6 disyllabic tokens and 11 monosyllabic tokens. All disyllabic tokens have the crucial segment in onset position. Eight of the words for each segment contain [ɪ] and are all monosyllabic (6 onset, 2 coda). The training tokens are controlled for frequency using the Subtlexus corpus Brysbaert and New (2009). Due to the limited number of words matching the criteria, there is a mismatch in frequencies. The [s]-words are more frequent (47.56/million) than the [f]-words (19.39/million), but no statistically significant difference is found between the log frequencies of the two groups [t(28.67) = -0.35, p = 0.73].

The new 34 words for the LDT were spoken by the same female native American English speaker from Michigan, as in Experiment 1. Each word was read aloud into a Logitech 980186-0403 microphone (frequency repsonse 100Hz–16kHz; -67dBV/ubar, -47dBV/Pascal ±-4dB) in a quiet room and recorded directly into Praat (Boersma and Weenink, 2016) at a sampling rate of 44.1 KHz. The speaker also recorded tokens of [fa], and [sa] to make the fa~sa continuum. Stimulus creation was done the same way as Experiment 1. The created fa~sa continuum was used in a new pre-test to find the most ambiguous step. Eight American English speakers (Mean Age = 20.5 years; 2 female) from Michigan State University separate from those in the main experiment participated in the pre-test for partial class credit. Participants heard the same 14 steps as in Experiment 1 (1,7,9,11,13,15,17,19,21,23,25,27,29,41) The mean response for each step was once again calculated to create the identification response function. and resulted in the expected sigmoidal function. Figure 4.1 shows these results below.

For the fa~sa continuum, step 16 was chosen as the most ambiguous segment. According to the pre-test results, the 50% response rate should be between steps 13 and 15, but due to the sharp rise and then drop between steps 13 and 17, multiple steps were tested against personal intuition in order to confirm that the segment chosen was maximally ambiguous. Steps 14 and 16 were tested due to their placement on the resulting categorical function from the pre-test. If we look at the area between steps 13 and 17 without step 15, then these both appear in a region that could feasibly be the 50% response rate. Step 18 was also tested as a comparable due to the result from the fi~si

continuum. Ultimately, step 16 was chosen to create the [$?_{fs-a}$] sound that would replace all of the [f] sounds in the LDT for the $FASA_{test}$ group.

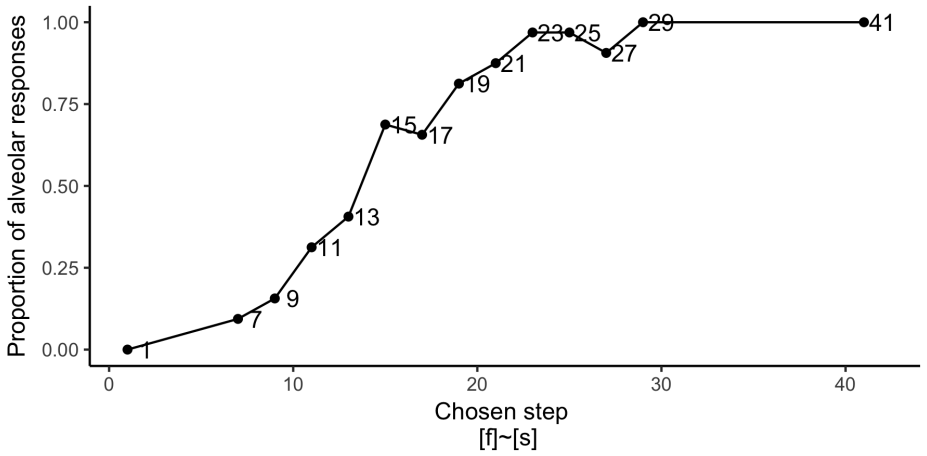

Figure 4.1: **Categorization results for the fa~sa continuum pre-test**

Praat (Boersma and Weenink, 2016) was once again used to create an altered versions of the LDT wordlist. The $FASA_{test}$ version of the list had all of the [f] tokens in the LDT replaced with [$?_{fs-a}$]. The $FISI_{control}$ wordlist was unaltered and therefore had no instances of an ambiguous token. To make the altered wordlist, the same method was used as Experiment 1. The resulting wordlist was then manually checked for naturalness based on author-intuition. While it is potentially worrisome since it has been shown that identification of the vowel following a fricative can be correctly made from acoustic information in the frication portion of the signal (Yeni-Komshian and Soli, 1981; Soli, 1981; McMurray and Jongman, 2016), previous experiments have used a similar method with no noticeable problem (Norris et al., 2003; Eisner and McQueen, 2005, 2006; McQueen et al., 2006a,b; Durvasula and Nelson, 2018).

4.1.4 Procedure

The general procedure is the same as outlined in Experiment 1 above. The few differences are outlined in the following paragraph. Participants were randomly assigned into one of two groups: $FASA_{test}$ or $FASA_{control}$. For the phonetic categorization tasks, both groups categorized the fa~sa

continuum. For the LDT, participants heard the list that corresponded to their assigned group as described above. Psychopy (Peirce et al., 2019) was once again used and all other procedural methods were exactly the same as Experiment 1.

4.2 Results

The same criteria as Experiment 1 was used to exclude any participants from analysis. Three participants failed to identify target segments accurately and were therefore removed. The $FISI_{control}$ group had an overall response accuracy rate of 91%, while the $FASA_{test}$ group's accuracy was lower at 87%.

A within-subjects analysis was again performed first and a one-tailed, paired t-test showed that there was once again a statistically significant decrease in alveolar responses from the "Before" phonetic categorization task to the "After" phonetic categorization task for the $FASA_{test}$ group [$\text{MeanDiff}_{\text{Before-After}}$ = -0.065, t(27)=-3.19, p<.01]. This suggests that participants in this group altered their categorization of the fa~sa continuum after hearing [$?_{fs-a}$] in the context of [i] and [ɪ] vowels in words that normally contained [f] during the lexical decision task. Unlike Experiment 1, the control group in this experiment did not show a shift towards the labiodental side of the continuum. A one-tailed, paired t-test showed that there was no decrease in alveolar responses from the "Before" phonetic categorization task to the "After" phonetic categorization task for the $FASA_{test}$ group [$\text{MeanDiff}_{\text{Before-After}}$ = -0.03, t(39)=-1.29, p=0.10]. The categorization functions for both of these comparisons can be seen in the upper left and upper right graphs of Figure 4.2.

Looking at the between-subjects comparisons, the standard "After"-"After" comparison for lexical retuning experiments resulted in no difference between the groups. A one-tailed Welch test showed no significant difference between the "After" responses of each group [$\text{MeanDiff}_{\text{Test-Control}}$ = -0.04, t(76)=-1.07, p=0.87]. The final comparison to be made is that of the difference between the two groups. This likely best captures the effect of training on the classification of a continuum as it has a baseline to compare against. It additionally allows for a normalization of the two groups. A one-tailed Welch test showed that there was no statistically significant difference in alveolar

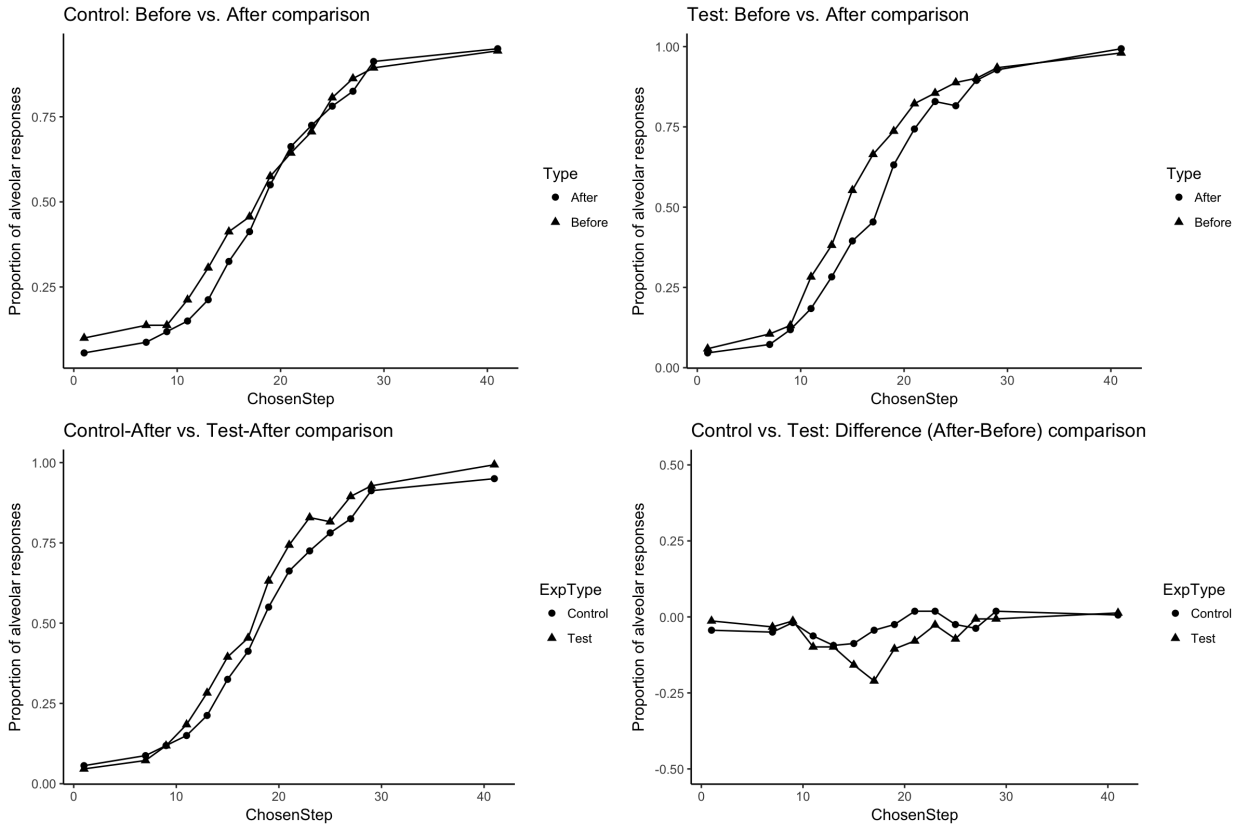

Figure 4.2: **Categorization results for Experiment 2**. Upper left is $FASA_{control}$ before vs after comparison; Upper right is $FASA_{test}$ before vs after comparison; Bottom left is the after responses of $FASA_{control}$ vs the after responses $FASA_{test}$; Bottom right is the difference between after and before responses between the two groups.

responses between the two groups when testing the entire continuum [$\text{MeanDiff}_{\text{Test-Control}} = 0.07$, $t(55.3)=2.34$, $p=0.011$]. Using the entire continuum for this analysis is rather conservative as there is no expected change at the tails. A priori, a decision was made to also run comparisons using the most ambiguous step and the two steps surrounding it (± 1).[1]. A one-tailed Welch test showed that there was a statistically significant difference in the difference in alveolar responses between the two groups when testing this subset of the continuum [$\text{MeanDiff}_{\text{Test-Control}} = 0.11$, $t(71.7)=1.97$, $p=0.026$]. This suggests that the $FASA_{test}$ group gave more labiodental responses after the lexical

[1]Recall that step 16 was determined to be the most ambiguous step in the pre-test and used as [$?_{fs-a}$]. Since step 16 was not used in the main experiment, a separate analysis on the "Before" responses of the $FASA_{control}$ group was used. This showed that step 17 was the most ambiguous, so steps 15,17,19 were used

decision task and therefore were able to generalize the learning effect from the context of high-front vowels to low-back vowels. These results are shown in the bottom left and bottom right graphs of Figure 4.2

4.3 Discussion

Experiment 2 shows that lexically-induced perceptual learning persisted under strict training and testing conditions. These results expand on those from Experiment 1 and show that generalization occurs even when the training and testing environments overlap. Listeners who heard [$?_{fs-a}$] in words that normally contained [f] during the LDT were more likely to respond "f" when categorizing a fa~s continuum afterwards. Unlike Experiment 1, the within-subjects comparisons differed from group to group. The $FASA_{test}$ group showed a significant difference in responses from before to after in the expected direction. In Experiment 1, both groups showed this difference and it was thought that there was a confounding effect, possibly from the formant transitions cues present in the ambiguous segment. The lack of difference here could be a result of using a different ambiguous token and continuum than the ones from Experiment 1. The vowel used in the fɑ~sɑ continuum may have weaker formant transition cues than the fi~si vowel. This is not to suggest that the type of vowel used ([i] or [ɑ]) is the locus of variation, but rather a byproduct of the two sets of stimuli being created at separate times. It is possible that during the construction process, the two vowel tokens were spliced at different time periods within the full [fi] or [fɑ] tokens such that they included different amounts of formant transition information. Ultimately, this variation between the control groups before-after responses from Experiment 1 to Experiment 2 is not super important as it gets washed out when looking at the difference of differences between the control and test groups.

The between group comparison showed the crucial difference. If we were to limit the analysis to just the 'After"-"After" comparison, as has previously been done in both lexical retuning and audio-visual recalibration experiments, then it would appear as if the effect is no longer able to generalize when we strictly control the training and testing conditions. This matches what was found in Reinisch et al.'s (2014) study and adds a new dimension to Van Linden and Vroomen's

(2007) story that these two paradigms tap into the same general perception mechanism. The comparison developed in Experiment 1 of the within group difference from before to after shows its importance here. Visually, it is very noticeable in the bottom right graph of Figure 4.2. There is clear separation in the before to after results for both groups and it once again happens in exactly the predicted section of the continuum.

This results is important because it not only has theoretical importance in regards to the generalizable nature of lexical retuning, but also methodological and analytical importance. The addition of the "Before" phonetic categorization in the experimental design allowed for better control of the sampling populations and subsequently a better analysis by comparing the amount of change between groups after training rather than just their responses after training. In the future when performing perceptual learning experiments, this comparison should allow for clearer insights into how listeners are changing their categorization habits in response to the training they receive. On the flip side, it is possible that previous experiments have missed certain insights by not making this comparison. Overall, the results from Experiment 2 confirm that lexically-induced perceptual learning (i.e., lexical retuning) is able to generalize to new contexts even when the training context is strictly reduced. Comparing this to the results presented in Reinisch et al. (2014), it suggests that lexical information is privileged in a way that auditory information is not (at least when it comes to perceptual learning). There is still an alternative explanation for the difference in results. The current lexical retuning experiment was designed in a way to provide type experience to listeners, while audio-visual recalibration experiments provide token experience. In order to fully comprehend whether the generalizable nature is domain specific or not, the type/token distinction needs to be further tested. Experiment 3 will do so.

CHAPTER 5

EXPERIMENT 3: REDUCING STIMULUS VARIATION IN LEXICAL RETUNING

One area that has not been addressed in the literature is the stimulus variation within each paradigm. Audio-visual recalibration experiments present the same string (typically some sort of VCV nonce word) continuously in order to induce the perceptual shift. Lexical retuning experiments present multiple, unique words throughout the LDT. In the former case, there is no variation, while the latter includes within-experiment stimulus variation. This experiment will test what happens if you remove the within-experiment stimulus variation from a lexical retuning experiment. Like Experiment 2, it will test whether training in the context of an [i] vowel will lead to retuning effects in a fa~sa continuum. The difference here is that now instead of 17 different training words, the LDT will contain 1 training word that gets repeated 17 times.

5.1 Method

5.1.1 Participants

Seventy-two undergraduate students from Michigan State University (Mean Age = 20.4; 54 female, 4 unreported gender) received either course credit or a small monetary reimbursement ($10) for participating in the study. All speakers identified as American English speakers and did not report any hearing problems.

5.1.2 Design

The overall design for Experiment 3 is identical to that in Experiment 2. Experiment 3 has the same three part set up of pre-training phonetic categorization, training *vis-à-vis* an auditory LDT, and post-training phonetic categorization. The only change made is to the LDT and will be outlined below. For this experiment, participants were assigned to one of two groups: $FASA2_{test}$ or $FASA2_{control}$.

5.1.3 Materials

The LDT for Experiment 3 uses a subset of the stimuli used in Experiment 2. The overall number of unique words used in the LDT for Experiment 3 is reduced to eight. Of the eight words, two are training items and the remaining six are filler. Two of the filler items are real English words and the other four are phonotactically licit English nonce words. For the training items, the [f]-word is more frequent (logFreq=3.21) than the [s]-word is (logFreq=2.64). Both training items are disyllabic and have the crucial segment in the onset position (word initial). Putting the training segment in onset position means there is a match for position between training and testing and therefore predicts a stronger likelihood that the retuning will occur. Using disyllabic words gives more information to the listener to identify it as a real English word and therefore gives a better chance that the ambiguous token will be recognized as a normal token of [f].

The filler items are all either mono- or di-syllabic. Two forms of the list were created. The $FASA2_{test}$ list was sampled from the *FASA* LDT list from Experiment 1 and therefore had the word containing [f] replaced with [$?_{fs-a}$]. The $FASA2_{control}$ list was identical to the $FASA2_{test}$ list, but it had the normal pronunciation for all words, including the word containing [f]. Therefore, the only difference between the two lists was in the single [f]-containing training token. The materials for the phonetic categorization tests are identical to the ones used in Experiment 2. Since the LDT materials are a subset of the ones created in Experiment 2, it was unnecessary to run a pre-test to find the most ambiguous segment to use for replacement in the LDT training words.

5.1.4 Procedure

The general procedure is the same as previous experiments. The few differences are outlined in the following paragraph. Participants were randomly assigned into one of two groups: $FASA2_{test}$ or $FASA2_{control}$. For the phonetic categorization tasks, both groups categorized the fa~sa continuum. For the LDT, participants heard the list that corresponded to their assigned group. Since the LDT lists for this experiment only contains 8 total words, each word was now presented 17 times each. 17 is chosen because that is the number of unique training words containing [f] or [s] in Experiment

1. Participants therefore heard 136 tokens in random running order during the lexical decision task. While the overall number of words participants heard was reduced by 14, they did hear the same number of training tokens between Experiment 2 and Experiment 3. Psychopy (Peirce et al., 2019) was once again used and all other procedural methods were the same as Experiments 1 and 2.

5.2 Results

The same criteria as in the previous experiments was used to exclude any participants from analysis. Five participants failed to identify target segments accurately and were therefore removed. Both groups had lower overall accuracy rates than the previous experiments (88% for $FASA2_{control}$and 85% for $FASA2_{test}$), but this was brought down primarily by the nonce words. For the training words containing the target fricatives, both groups had greater than 95% accuracy.

Both groups showed no difference in alveolar responses from "Before" to "After". A one-tailed, paired t-test showed that there was no statistically significant decrease in alveolar responses from the "Before" phonetic categorization task to the "After" phonetic categorization task for the $FASA2_{test}$ group [$\text{MeanDiff}_{\text{Before-After}}$ = -0.032, t(31)=-1.25, p=0.11] and for the $FASA2_{control}$ group [$\text{MeanDiff}_{\text{Before-After}}$ = -0.013, t(34)=-1.21, p=0.12]. This suggests that in both cases there was no observable change in alveolar responses after the lexical decision task for either group. What is of note here is the fact that the $FASA2_{test}$ group did not show a statistically significant change despite hearing the same overall number of [$?_{fs-a}$] tokens (17) as the test group in Experiment 2. The categorization functions for both of these comparisons can be seen in the upper left and upper right graphs of Figure 5.1.

Looking at the between-subjects comparisons, the 'After"-"After" results do show a shift in the expected direction, but not enough to be statistically different between the groups. A one-tailed Welch test showed that there was no statistically significant difference in the percentage of alveolar responses between the $FASA2_{test}$ and $FASA2_{control}$ groups in the "After" phonetic categorization task [$\text{MeanDiff}_{\text{Test-Control}}$ = 0.064, t(61.5)=1.47, p=0.07]. As discussed in previous chapters, the difference comparison is likely the better comparison to make. A one-tailed Welch test showed that

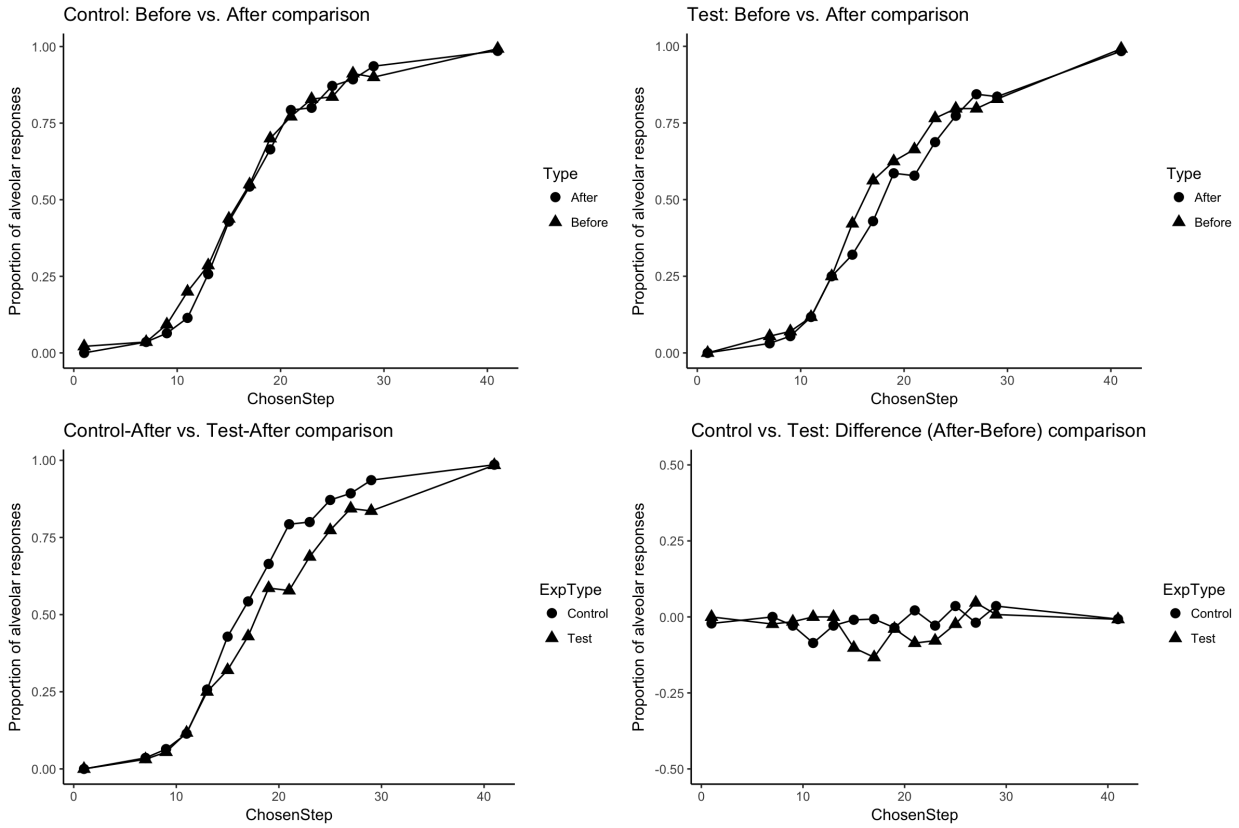

Figure 5.1: **Categorization results for Experiment 3**. Upper left is $FASA2_{control}$ before vs after comparison; Upper right is $FASA2_{test}$ before vs after comparison; Bottom left is the after responses of $FASA2_{control}$ vs the after responses $FASA2_{test}$; Bottom right is the difference between after and before responses between the two groups.

there was no statistically significant difference in the difference of alveolar responses from "Before" to "After" between the two groups [$\text{MeanDiff}_{\text{Test-Control}} = 0.02$, $t(41.1)=0.702$, $p=0.24$]. This result seems to corroborate the idea that the lexical retuning effect did not occur for the participants in the $FASA2_{test}$ group. The between-subjects results are shown in the bottom left and bottom right graphs of Figure 5.1.

5.3 Discussion

The results from Experiment 3 question whether the locus of variation in the generalizability of perceptual learning effects is the modality in which listeners receive their disambiguating information. It was hypothesized that because lexical information is considered to be high-level

information, that it would allow for generalization due to tapping into the representational structure of the segments. The low-level visual information not tapping into the representational structure was subsequently proposed for why Reinisch et al. (2014) observed no generalization effects using an audio-visual recalibration paradigm. This experiment brings into question that hypothesis.

In Experiment 2, it was observed that listeners could generalize to a new context in the testing condition when the training condition contained one unique vowel environment presented in multiple unique words. Experiment 3 kept the same testing condition and the same single vowel environment for the training condition, but repeated the same word multiple times. In other words, it traded type experience for token experience to better match the conditions used in audio-visual recalibration experiments. The results suggest listeners were no longer able to generalize the lexical retuning effect, despite hearing the same number of raw tokens in the training condition for both experiments. This suggests a possible reanalysis of why Reinisch et al. (2014) were unable to find any type of generalization effects. It may not be the case that visually-guided perceptual learning is specific to phonemic contrast, cues, and contexts as they claim (Reinisch et al., 2014, p. 102), but rather that generalization of the effect requires type experience. Because the experiments were designed in a way that did not allow for type experience, it is now unclear as to what is the exact cause of the lack of generalization.

That is not to say that there still may not be a difference in generalizability between the two paradigms. What this experiment has shown, in conjunction with the previous experiments, is that type experience is necessary to learn and generalize the perceptual learning effect in lexical retuning experiments. This may just be a general property of learning linguistics generalizations as it has also shown to be beneficial in teaching infants stress constraints (Gerken and Bollt, 2008) as well as in phonotactic learning experiments (Denby et al., 2018). It is therefore necessary to see how type experience affects generalization in audio-visual recalibration experiments in order to confirm that the difference between it and the lexical retuning paradigm can be explained by this general learning principle. Future experiments beyond the scope of this thesis may be beneficial in this regard.

There are two potential outcomes for this hypothetical experiment. On one hand, if an audiovisual-recalibration experiment is designed to allow for type variation in the training stimuli, and no generalization to new environments is found, this would suggest that it is the modality of the disambiguating information that is the cause of the previously observed variation in results. In other words, lexical information leads to generalization while visual information does not. If generalization is observed in this new experiment, it would suggest that the two disambiguating modalities are being treated equally and the previously observed differences are likely due to the type/token distinction in the training stimuli. Regardless of the outcome, the results should set an important precedence for further experiments using either of the perceptual learning paradigms.

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 Summary of Results

This thesis investigated the generalizable nature of perceptual learning in speech using the lexical retuning paradigm developed by Norris et al. (2003). More specifically, it sought out why the two paradigms housed under the perceptual learning tent (lexical retuning and audio-visual recalibration (Bertelson et al., 2003)) were showing different results in relation to generalization despite earlier claims that both were part of the same general perception mechanism (Van Linden and Vroomen, 2007). Three new experiments were run to confirm that lexical retuning allowed for generalization to new contexts both under broad (Experiment 1) and more restricted (Experiment 2) training conditions, as well as what effect type and token experience have on generalization (Experiment 3).

The results from Experiment 1 showed that generalization of the perceptual learning effect in lexical retuning experiments was possible when using a multitude of different training environments and testing on an environment that listeners only heard once throughout the training phase. Listeners were more likely to respond "f" on a phonetic categorization task of a fi~si continuum after hearing an ambiguous segment [?$_{fs-i}$] in place of [f] in words during a lexical decision task than listeners who heard normal [f] in the same words. Experiment 1 also presented new methodological and analytical results for perceptual learning studies. Previously, comparisons of phonetic categorization responses after exposure to an ambiguous segment within both paradigms have been made between two groups. This tacitly assumes a homogenized beginning state for all individuals. The addition of a "Before" phonetic categorization task to the experiment design allowed for a comparison of differences from before to after between the two groups. In Experiment 1, it was this comparison of the differences that most explicitly displayed the effect that the lexically-guided training provided by the ambiguous segment had on the test group. Experiment 1 was therefore

important not only to corroborate past results and give more supporting evidence for generalization within lexical retuning, but also to show that methodologically there are places in which we can improve the experimental/analytical methods used within these perceptual learning paradigms.

Experiment 2 observed that lexical retuning could still lead to generalization, even in more strictly controlled training and testing conditions. Unlike Experiment 1 where the set of training words contained multiple unique vowel environments, Experiment 2's training set only included words that contained either an [i] or an [ɪ] vowel. Listeners were then tested on a fa~s continuum. Once again, listeners were more likely to respond "f" on a phonetic categorization task the fa~sa continuum after hearing an ambiguous token [$?_{fs-a}$] in place of [f] in words during a lexical decision task than listeners who heard normal [f] in the same words. This result was especially clear when looking at the difference from "Before"-"After" between the test and the control group. The importance of this comparison was also made clear with Experiment 2. The 'After"-"After" comparison between the two groups showed no statistically significant difference and in fact showed more alveolar responses for the test group (opposite of what is predicted). Overall, the results from experiment challenge what was shown with audio-visual recalibration experiments, mainly that the phonological environment needs to be identical in training and testing conditions in order to see an effect (Reinisch et al., 2014).

Experiment 3 was used to test whether or not type experience could explain why lexical retuning experiments allowed for generalizability. The training and testing environments for Experiment 3 were identical to the ones used in Experiment 2, but the overall number of training items was reduced. In Experiment 2, listeners heard 34 unique training words, 1 time each (17 containing [$?_{fs-a}$]; 17 containing [s]). In Experiment 3, listeners heard 2 unique training words, 17 times each (1 containing [$?_{fs-a}$], 1 containing [s]). This was done to attempt to match the training experience that participants typically receive in audio-visual recalibration experiments. In those experiments, the same training stimuli is presented over and over again. Results from Experiment 3 show that listeners were no longer able to generalize their perceptual training when presented with the same ambiguous stimuli multiple times rather than multiple, unique stimuli. In this case, there

was no statistically significant difference between the test and control groups and suggests that type experience is necessary for generalization within lexical retuning.

6.2 General Discussion

One idea presented early on in this thesis was to probe whether or not different types of disambiguating information were treated differently by the speech perception mechanism. Based on the results presented here, I can not conclusively answer this question. What has been shown is that when the training and testing conditions are more closely matched for the two perceptual learning experimental paradigms, both unambiguous visual and unambiguous lexical information do appear to show similar results, specifically in the realm of generalizability. The results from Experiment 3 most explicitly showed that removing type variability in the training set resulted in a lack of generalizability within the lexical retuning paradigm, just as with the audio-visual recalibration paradigm. This suggests that it would be hasty to attribute previously observed differences in generalizability between the two paradigms to different underlying mechanisms. However, it should be pointed out that in order to complete the argument, it is now crucial to see what happens when the audio-visual recalibration paradigm is altered to contain type variation as in Experiments 1 and 2. I leave this as a demonstration for further research.

In general, the ability for generalization within the lexical retuning paradigm supports a view of speech perception involving abstract representations (i.e., phonemes, features). This view has been questioned by more recent experimental results (Reinisch et al., 2014; Mitterer et al., 2016, 2018), but the results presented in this thesis put them into question. The generalizable nature of perceptual learning has up to this point been almost exclusive to the lexical retuning paradigm, where it has been shown to generalize over position (Jesse and McQueen, 2011), features (**?**Durvasula and Nelson, 2018), and to new words (McQueen et al., 2006a; Sjerps and McQueen, 2010). Reinisch et al.'s (2014) provide the strongest argument against generalization using the audio-visual recalibration paradigm, but this thesis has identified a critical difference between the two experiment styles that may have lead to the contradictory claims. In this regard, the lack of generalization shown in

Reinisch et al. (2014) could be taken as a result of no type experience in the training stimuli and not assumed to be a result of the lack of abstract representations within speech perception.

6.3 Conclusion

An important question within speech perception research is what do listeners do when presented with ambiguous input? This thesis has shown that the answer to that is largely dependent on the type of disambiguating information they receive. What is most clear is that there appears to be a type/token distinction. When listeners hear ambiguity in only one specific type, they consider it to be an idiosyncrasy and do not generalize it to other types, even when hearing multiple repetitions of the token. This is an alternative explanation to previous perceptual learning results in the audio-visual recalibration domain (Reinisch et al., 2014) that claimed that the effect was highly constrained to conditions that completely matched the testing environment. The necessity of type variation has also been shown in other learning domains (Gerken and Bollt, 2008; Denby et al., 2018) and suggests that this may not entirely be specific to perceptual learning paradigms. It is therefore necessary to see how type experience interacts with visually-guided perceptual learning in order to tease apart the roles that different modalities have on the speech perception mechanism.

Overall this thesis has made three contributions. First, it has corroborated the idea that lexical retuning can result in generalization. This was observed under broad conditions in Experiment 1 and more strict conditions in Experiment 2. Second, it has shown that type experience is crucial for the generalization to occur. Third, it has argued for a change to the methodology and analysis of perceptual learning experiments by adding the "Before" phonetic categorization task to the experimental design. This subsequently allowed for better insight into how the ambiguous token presented in the lexical decision task affects the categorical representation of the segments under question by allowing for a comparison of the differences within test and control groups.

APPENDICES

APPENDIX A

LEXICAL DECISION TASK WORD LIST

Training Words			
Experiment 1		Experiment 2/3	
bluff	truss	cliff	kiss
chef	chess	reef	geese
cliff	bliss	thief	piece
deaf	less	whiff	bliss
poof	deuce	film	silk
whiff	kiss	filth	sick
beef	geese	fish	sim
cough	boss	fill	sing
fudges	such	fifth	sip
felt	sect	fib	seek
food	soup	fiend	seem
fab	sash	feeble	seagull
fade	say	**female**	**seated**
fig	sip	fetal	seeing
fish	silt	fever	seeker
fool	soon	fiji	seeping
full	sulk	feline	seething

Table A.1: **Training words for Lexical Decision Tasks**. The bolded subset indicate Experiment 3.

Filler Words					
Real		Nonce			
truck	rowboat	**himp**	**plone**	heen	**weg**
tomb	tuba	rone	twek	hoong	yilk
right	laker	relnt	trag	gelp	queng
loot	daydream	trad	toin	roip	yoog
wet	butter	waip	rone	giblo	ledmon
trim	little	ligbee	poindool	wegring	troplund
pen	data	yoomake	metroy	talkibd	yoogul
lick	liquor	rekmate	deemlond	retlid	**cratbool**
mole	maker	magmilnt	ratwell	bidbowk	pingloon
mall	treaty	mitger	petnole	lempok	kotlim
twin	trooper	wiblee	hapekt	libmadoop	pendalimt
lake	toilet	midaran	wegtamlip	kebnoti	cridamlo
line	lemur	lopenad	dripnalog	yoodalin	
rat	baking	doonagort	himnapile	drataloob	
while	typing	rivathrog	podaling	datoonmeg	
will	boycott	noopalingdo	habapithloo	boondatribnok	
day	pillow	poginito	habalipdee	hundapegnort	
drip	yellow	dewagelin	blegondanit	kililitbi	
globe	jupiter	krebdingolee	bindilamok	kinudiplo	
toy	lump	tridalondit	radoobling	kablingno	
building		plagtorin	trongolemp	hapooli	

Table A.2: **Filler words for Lexical Decision Tasks**. All 116 words were used in Experiments 1 and 2. The bolded subset were used in Experiment 3

APPENDIX B

LDT TRAINING WORD FREQUENCY INFORMATION

B.1 Experiment 1

SYLLABLES	POSITION	WORD	FREQ/MIL	LOGFREQ	WORD	FREQ/MIL	LOGFREQ
1	coda	bluff	6.22	2.5	truss	0.33	1.26
1	coda	chef	11.88	2.78	chess	7.45	2.58
1	coda	cliff	21.57	3.04	bliss	3.14	2.21
1	coda	deaf	14.53	2.87	less	111.1	3.75
1	coda	poof	2.16	2.05	deuce	2.86	2.17
1	coda	whiff	2.49	2.11	kiss	121.16	3.79
1	coda	beef	19.71	3	geese	1.59	1.91
1	coda	cough	8.78	2.65	boss	124.29	3.8
1	onset	fudge	3.63	2.27	such	291.22	4.17
1	onset	felt	119.82	3.79	sect	0.73	1.58
1	onset	food	154.43	3.9	soup	25.2	3.11
1	onset	fab	0.61	1.51	sash	1.14	1.77
1	onset	fade	5.61	2.46	say	1639.78	4.92
1	onset	fig	1.22	1.8	sip	5.1	2.42
1	onset	fish	83.49	3.63	silt	0.33	1.26
1	onset	fool	89.33	3.66	soon	257.65	4.12
1	onset	full	166.9	3.93	sulk	0.75	1.59

Table B.1: **Experiment 1 LDT training word frequency data**

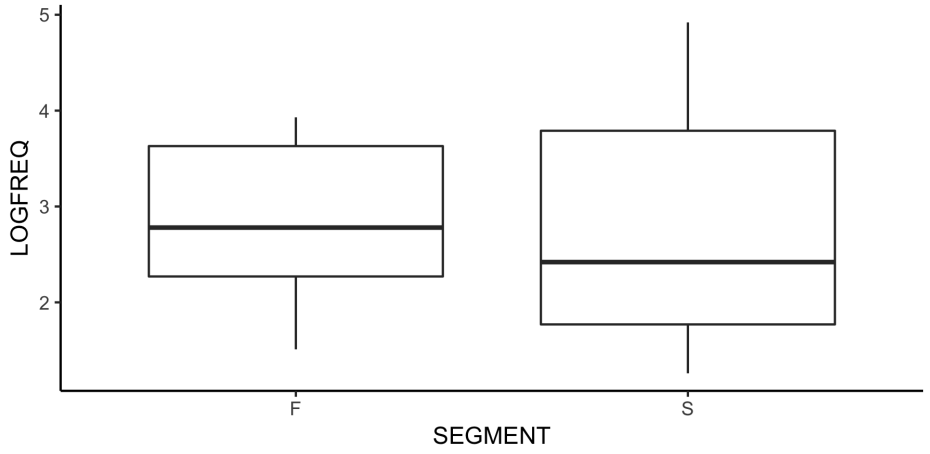

Figure B.1: **Boxplots for Experiment 1 LDT training word frequency (log)**

B.2 Experiments 2/3

SYLLABLES	POSITION	WORD	FREQ/MIL	LOGFREQ	WORD	FREQ/MIL	LOGFREQ
1	coda	cliff	21.57	3.04	kiss	121.16	3.79
1	coda	reef	4	2.31	geese	1.59	1.91
1	coda	thief	24.27	3.09	piece	124.49	3.8
1	coda	whiff	2.49	2.11	bliss	3.14	2.21
1	onset	film	65.25	3.52	silk	9.78	2.7
1	onset	filth	4.53	2.37	sick	165.43	3.93
1	onset	fish	83.49	3.63	sim	1.1	1.76
1	onset	fill	43.94	3.35	sing	97.59	3.7
1	onset	fifth	19.2	2.99	sip	5.1	2.42
1	onset	fib	1.27	1.82	seek	18.31	2.97
1	onset	fiend	2.8	2.16	seem	139.82	3.85
2	onset	feeble	1.69	1.94	seagull	1.22	1.8
2	**onset**	**female**	**31.61**	**3.21**	**seated**	**8.55**	**2.64**
2	onset	fetal	1.16	1.78	seeing	109.63	3.75
2	onset	fever	19.94	3.01	seeker	0.92	1.68
2	onset	fiji	1.47	1.88	seeping	0.37	1.3
2	onset	feline	0.9	1.67	seething	0.45	1.38

Table B.2: **Experiment 2/3 LDT training word frequency data**. All words were used in Experiment 2. The bolded subset were used in Experiment 3.

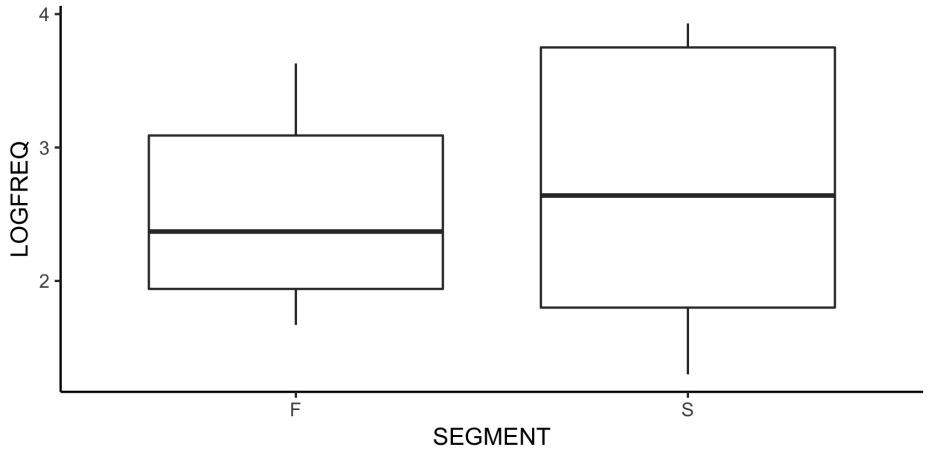

Figure B.2: **Boxplots for Experiment 2 LDT training word frequency (log)**

BIBLIOGRAPHY

BIBLIOGRAPHY

Baart, M. and Vroomen, J. (2010). Phonetic recalibration does not depend on working memory. *Experimental brain research*, 203(3):575–582.

Bertelson, P., Vroomen, J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science*, 14(6):592–597.

Boersma, P. and Weenink, D. (2016). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/.

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Cutler, A., McQueen, J. M., Butterfield, S., and Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boudaries.

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

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1962). Formant transitions and loci as acoustic correlates of place of articulation in american fricatives. *Studia linguistica*, 16(1-2):104–122.

Denby, T., Schecter, J., Arn, S., Dimov, S., and Goldrick, M. (2018). Contextual variability and exemplar strength in phonotactic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(2):280.

Durvasula, K., Huang, H.-H., Uehara, S., Luo, Q., and Lin, Y.-H. (2018). Phonology modulates the illusory vowels in perceptual illusions: Evidence from mandarin and english. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*.

Durvasula, K. and Kahng, J. (2015). Illusory vowels in perceptual epenthesis: The role of phonological alternations. *Phonology*, 32(3):385–416.

Durvasula, K. and Kahng, J. (2016). The role of phrasal phonology in speech perception: What perceptual epenthesis shows us. *Journal of Phonetics*, 54:15–34.

Durvasula, K. and Nelson, S. (2018). Lexical retuning targets features. In *Proceedings of the Annual Meetings on Phonology*, volume 5.

Eimas, P. D. and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1):99 – 109.

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

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

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, 6(1):110.

Gaskell, M. G. and Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human perception and performance*, 22(1):144.

Gaskell, M. G. and Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2):380.

Gerken, L. and Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, 4(3):228–248.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, 65(4):575–590.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111.

Jesse, A. and McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, 18(5):943–950.

Kleinschmidt, D. F. and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148.

Kraljic, T. and Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, 51(2):141–178.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

McMurray, B. and Jongman, A. (2016). What comes after/f/? prediction in speech derives from data-explanatory processes. *Psychological science*, 27(1):43–52.

McQueen, J. M., Cutler, A., and Norris, D. (2006a). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30:1113–1126.

McQueen, J. M., Norris, D., and Cutler, A. (2006b). The dynamic nature of speech perception. *Language and speech*, 49(1):101–112.

Mitterer, H., Chen, Y., and Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, 35(1):184–197.

Mitterer, H., Cho, T., and Kim, S. (2016). What are the letters of speech? testing the role of phonological specification and phonetic similarity in perceptual learning. *Journal of Phonetics*, 56:110–123.

Mitterer, H., Reinisch, E., and McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, 98:77 – 92.

Mitterer, H., Scharenborg, O., and McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, 129(2):356–361.

Norris, D., McQueen, J. M., and Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3):299–325.

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 30(2):1113–1126.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, pages 1–9.

Reinisch, E. and Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2):539.

Reinisch, E., Wozny, D. R., Mitterer, H., and Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of phonetics*, 45:91–105.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive psychology*, 18(4):452–499.

Samuel, A. G. and Kat, D. (1998). Adaptation is automatic. *Perception & psychophysics*, 60(3):503–510.

Samuel, A. G. and Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6):1207–1218.

Scharenborg, O., Mitterer, H., and McQueen, J. M. (2011). Perceptual learning of liquids. In *Interspeech 2011: 12th Annual Conference of the International Speech Communication Association*, pages 149–152.

Sjerps, M. J. and McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 36(1):195.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, 70(4):976–984.

Van Linden, S. and Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6):1483.

Vroomen, J., van Linden, S., Keetels, M., De Gelder, B., and Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, 44(1-4):55–61.

Yeni-Komshian, G. H. and Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, 70(4):966–975.