Title
Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent

Author
Melguy, Yevgeniy V

Publication Date
2022-10-01

Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent

by

Yevgeniy Vasilyevich Melguy

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Keith Johnson, Chair
Professor Frederic Theunissen
Professor Terry Regier
Assistant Professor Isaac Bleaman

Fall 2022

Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent

Abstract

Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent

by

Yevgeniy Vasilyevich Melguy

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Keith Johnson, Chair

The aim of this dissertation is to investigate the mechanisms involved in phonetic learning of an unfamiliar accent, focusing on understanding what processes underlie changes to phonetic category structure, how such learning affects subsequent online lexical processing, and whether the same mechanisms that underlie learning for a single speaker are also responsible for generalization of learning to novel speakers with a similar pronunciation.

The first set of experiments (chapter 2) investigates the mechanisms that underpin the changes in phonetic category structure (lexically-guided phonetic recalibration), following exposure to a novel artificial accent. This chapter focuses on two possible adaptive strategies that listeners may use which have been suggested in the literature: phonetic category shift and phonetic category expansion. Under the first hypothesis, listeners make targeted adjustments to a category boundary based on the specific accent they encounter, whereas under the second, they utilize a more general expansion of that category in perceptual space. Results of two experiments suggest that listeners rely on a (nonuniform) category expansion strategy that is constrained by acoustic similarity to sounds involved in the exposure accent.

The experiments in Chapter 3 focus on the relationship between changes in category structure and more online measures of speech processing. It is often assumed in the literature on perceptual learning for speech that changes in phonetic category boundaries (recalibration) following exposure to an atypical pronunciation underlie improved comprehension and/or processing of accented speech. The experiments in this chapter test whether exposure to such an artificial accent can facilitate subsequent processing of accented words, and whether the same mechanisms that constrain category boundary changes found in Chapter 1 also obtain for lexical processing. Results suggest that accent exposure does result in changes to lexical processing, and results provide tentative support for a form of category expansion as the mechanism for such changes.

The last set of experiments (Chapter 4) examine how perceptual learning may generalize

to novel speakers and novel sound contrasts. The goal of this set of experiments is to test whether the same mechanisms that are responsible for category boundary changes in a single speaker are also applicable to novel speakers. Previous literature suggests that phonetic learning of certain speech sounds (e.g., fricatives) may be speaker-specific — it does not transfer to a novel speaker, possibly because fricatives contain spectral properties that cue speaker identity. However, the results of Chapter 1 indicate that transfer to a novel sound contrast can occur within a single speaker, suggesting that there is room for transfer of learning to a novel speaker even if their pronunciation is acoustically distinct from that of the exposure speaker. Results of these experiments show transfer of phonetic learning to both novel speakers and novel phonetic contrasts. While results are mixed, depending on speaker and contrast, there is tentative evidence that listeners may use a more targeted mechanism when generalizing learning to novel speakers.

Together these experiments indicate that lexically-guided phonetic learning is flexible enough to accommodate differences between familiar and novel contexts, suggesting that it may be a viable mechanism for adapting to variability in speech. However, while results provide evidence for some degree of generality in the underlying perceptual learning mechanisms, they also show that such generalization is constrained by acoustic similarity to previous experiences, whether the latter involves novel sound contrasts, novel speakers, or both. This supports a view of the perceptual system as one that is dynamic but that must still balance stability and plasticity.

To my parents, Anna and Vasiliy.

Contents

List of Figures

List of Tables

Acknowledgments

First, thank you to all the members of my dissertation committee—Keith Johnson, Terry Regier, Isaac Bleaman, and Frederic Theunissen—for taking the time and effort to help guide me through this project, from its inception as a fuzzy set of ideas over two years ago to its conclusion today.

I am especially grateful to Keith for his mentorship throughout my doctoral career. Over six years ago I made the decision to come to Berkeley to work with Keith, and I am glad I chose to trust my gut on this one! Keith has always been supportive, kind, and patient with me. Our meetings were always a highlight of my week. Keith never made me feel dumb for asking a dumb question or being confused about something, and I always left his office feeling better about whatever problem I had been dealing with all week, whether personal or academic. He also tolerated all my wacky experiment ideas (only noting after the fact that "I didn't think this would work, but I figured I would let you try it!"). Keith was happy to provide assistancd when I needed it, but he also pushed me to be the independent researcher I am today, and taught me almost everything I know about phonetics!

Thanks to all of my colleagues and professors in Berkeley linguistics, for everything you have taught me over the course of my graduate career, and for forming my academic community for so many years. Special thanks is also due to the Berkeley linguistics staff for their tireless support. Ron Sprouse has always been a key source of technical help, and has been instrumental to the success of countless experiments. A huge thank you to Belén Flores, Paula Floro, and Johnny Morales Arellano for their administrative assistance, for making sure my experiments got funded, and for promptly responding to my panicked emails about degree completion deadlines! I would also like to thank my research apprentices, Andres Sanchez and Kenny Ho, for their assistance with various aspects of this project, from stimulus creation to debugging code. And of course, thank you to everyone who gave me permission to record (and subsequently manipulate!) their speech to produce the materials for this project. Your voices made this project possible.

My love and gratitude go to my amazing Berkeley Student Co-op family for the years of community, friendship, and affordable rent. Thank you for helping me weather the hardest moments of graduate school, and sharing my joy during the best moments. I am eternally grateful for all of the amazing relationships that I have been blessed with during my years in the co-ops. Thank you for the homemade pizza nights, the breakfast banter, the crazy construction projects, the dining room conversations about mind-body dualism, and all of the beautiful ordinary day-to-day moments that come with living in the same house with 40+ people. I really couldn't have done it without you. You know who you are.

I would also like to thank my parents, Anna and Vasiliy, for their support through 10 years of university life and 30 years of general life. I hope that I make you proud as the first doctor in the family (you can keep calling me Zhenya!). Last but not least, thank you to my partner Elena for the unconditional love and support in the long, hard final stretches of graduate school. It has been a comfort and a joy to have you by my side.

Chapter 1

Perceptual learning for speech

1.1 Introduction

Listeners must deal with a wide range of variability in speech, both within and across speakers. This is a non-trivial task, as many factors affect how a given speech sound can be produced . These include linguistic factors such as intonation, stress, and surrounding context. They also include anatomical characteristics (e.g., vocal tract length, palate shape), and social variables (e.g., race, gender, sexuality, socio-economic class, etc.).

The general goal of this dissertation is to investigate how listeners adapt to such speech variability, focusing on one factor that has been shown to pose particular trouble for listeners: a non-native accent or dialect. It is well known that an unfamiliar accent can result in comprehension difficulty for listeners. Studies have demonstrated that processing foreign-accented speech is more difficult for listeners, as shown by increased processing time (Clarke & Garrett, 2004; Munro & Derwing, 1995), lower word identification scores (Bent & Holt, 2013), and more exposure time required for word recognition (Leikin et al., 2009). Fortunately, although L2 learners of a language often do not attain native-like phonetic proficiency (Flege et al., 1995), listeners show a remarkable ability to rapidly adapt to accented speech (see Cristia et al. (2012) for a review). Such adaptation is a form of perceptual learning — a general phenomenon where experience with a stimulus leads to improved perception of that stimulus. An important question is how listeners are able to strike the right balance between specificity and generalization in perceptual learning for unfamiliar speech. What mechanisms allow listeners to adapt to a given speaker's accent, and what factors allow such learning to transfer to new contexts? For learning to be useful, it must be broad enough in scope that it can generalize to novel contexts. Learning an accent must mean, on the one hand, that the listeners can become attuned to the phonetic patterns that characterize that specific accent, and subsequently can generalize such learning to novel speakers with a similar pronunciation. On the other hand, it also means that listeners must be able to constrain the transfer of learning so that it does not interfere with processing of other speakers who do not share those phonetic characteristics.

In the context of accented speech, perceptual learning has frequently been investigated using either comprehension-based measures (Baese-Berk et al., 2013; Bradlow & Bent, 2008; Gordon-Salant et al., 2010; Melguy & Johnson, 2021; Sidaras et al., 2009; Vaughn, 2019), or measures of processing fluency (Clarke & Garrett, 2004; Xie, Weatherholtz, et al., 2018). Such studies typically evaluate the success of learning measures through measures such as sentence transcription or response latencies in some form of a word recognition task. Another important body of literature has investigated perceptual learning for speech with a focus on phonetic category structure. Perceptual learning in these cases takes the form of 'phonetic recalibration'[1], a process that involves listeners adjusting their perceptual category boundaries for a given speech sound following exposure to a speaker with an atypical pronunciation of that sound (Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006, 2007; Norris et al., 2003). Although these studies generally involve artificial 'accents' created by manipulating individual target phonemes, a common assumption is that such lexically-guided retuning of phoneme categories underlies listeners' ability to adapt to novel accents or dialects (Bradlow & Bent, 2008; Eisner et al., 2013; Witteman et al., 2013).

Phonetic recalibration offers a promising approach to investigating the mechanisms that underlie perceptual learning of accented speech, because the targeted manipulation of just a single phoneme and evaluation of subsequent changes in processing of that sound give us a fine-grained measure of how the perceptual system adjusts to accommodate an atypical pronunciation. This contrasts with the 'black box' approach of studies that use natural foreign accents — comprehension and processing-based measures are often too coarse-grained to shed much light on the mechanisms responsible for adaptation, and the multiple parameters of deviation from natively-accented speech often leave it unclear *what* listeners are adapting to. Crucially, while it has often been assumed that recalibration of phonetic categories facilitates comprehension and/or processing of accented speech (Bradlow & Bent, 2008; Eisner et al., 2013; Kraljic & Samuel, 2006), there remains no direct evidence for this assumption. An important goal of this dissertation is thus to investigate how recalibration (sound-specific adjustment of perceptual category boundaries) may benefit subsequent processing of accented speech. Do the same mechanisms that underlie recalibration or phonetic re-tuning of internal sound category structure also underlie accent accommodation more generally?

Before diving deeper into the research on perceptual learning for speech, this chapter briefly reviews the general literature on perceptual learning. Placing the current project in against the backdrop of perceptual learning across sensory domains may give the reader a better understanding of which aspects of learning may be restricted to the domain of language and which are more general principles that guide perceptual learning. The remainder of this chapter is organized as follows. Section 1.2 gives a brief overview of the general literature

[1]In the literature on perceptual learning for speech, the terms "phonetic recalibration," "phonetic retuning," "phonetic learning," and "perceptual learning" are often used interchangeably to refer to this phenomenon. In this study, we reserve the term "perceptual learning" to refer to the general phenomenon (across sensory domains), the term "phonetic learning" to refer to perceptual learning for speech, and the terms "phonetic recalibration," "recalibration," or "retuning" to refer specifically to the adjustment of perceptual category boundaries following exposure to an atypical pronunciation.

on perceptual learning, including psychophysical and neurological correlates of perceptual learning across domains and suggested mechanisms. Section 1.3 takes a closer look at the literature on perceptual learning for speech and how these findings may bear on the specific case of perceptual adaptation to accented speech. Section 1.4 concludes the chapter by providing an overview of the open questions in perceptual learning for accented speech that will be the focus of this dissertation.

1.2 Perceptual learning

What is perceptual learning?

Definitions of perceptual learning focus predominantly on perceivers' adaptation to stimulus properties which affect subsequent perception of those and/or similar stimuli. For instance, Fahle (2001) emphasizes low-level sensory adaptation as the key distinguishing feature: "Perceptual learning […] is a form of learning leading to better use of sensory information which is relatively independent of conscious or declarative forms of learning but relies partly on rather low-level modifications in the central nervous system. Perceptual learning hence resembles, in many respects, procedural forms of learning that are common in motor learning, for example learning to ride a bicycle." (p. 11225). Hall (2008) focuses on the acquired ability to attend more effectively to unique elements of similar objects while filtering out their shared properties, defining perceptual learning as "the learning process (or processes) that increases the effectiveness of unique stimulus elements and/or reduces that of common stimulus elements, thus facilitating discrimination between similar stimuli. (p. 110). Finally, Goldstone (1998) stresses the biologically adaptive elements of perceptual learning as "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment." (p. 1).

While these definitions of perceptual learning vary in their emphasis, a point of general consensus in the literature is that perceptual learning is not a simple practice effect that emerges from participant familiarization with task rules or strategies, but rather reflects substantial changes in sensory processing. Gold and Watanabe (2010) point to several pieces of evidence in favor of this idea. First, perceptual learning increases perceptual sensitivity, as measured by the lower quality of stimulus that is needed to reach a particular threshold of perceiver response. Second, perceptual learning effects are often highly specific to characteristics of the exposure stimuli and task. This failure of transfer to similar tasks indicates that perceptual learning cannot be solely attributed to familiarity with task procedures. Third, as other reviewers have noted, perceptual learning effects tend to be relatively long-lasting, often persisting over weeks or months (Fahle, 2005; Goldstone, 1998) which distinguishes them from other learning effects such as habituation, sensitization, or priming, which are generally short-term (Fahle, 2001; Gold & Watanabe, 2010).

Perceptual learning across sensory domains

While interest in perceptual learning for speech is a relatively recent phenomenon, perceptual learning has been a topic of investigation in experimental psychology since at least the mid-nineteenth century and has been attested across all sensory domains in a variety of simple and complex tasks.

One of the earliest studies documenting perceptual learning coincides with the birth of psychophysics — the study of the relation between physical stimuli and the perceptions they produce. Volkmann (1858) investigated the effects of practice on discrimination of tactile difference thresholds. For each trial, subjects received two consecutive pricks on a region of the arm and were asked to indicate whether the two points occurred in an identical spot or distinct spots of the arm. Over the course of the experiment, subjects became capable of detecting smaller and smaller distances between points, reducing the distance by more than half after only a few hours of training. Much of the research on perceptual learning has occurred in the visual modality, spurred by a seminal study by Gibson and Gibson (1955), who used an image identification task to assess perceptual learning. They first presented subjects with the target (a simple line drawing) and then with a succession of similar line drawings, asking them to indicate which were identical to the target. Over the course of the experiment, all participants showed a decrease in the number of incorrect responses and improved ability to specify the dimensions on which non-identical items differed from the target (e.g., too short, wrong orientation, etc.). The authors' interpretation of this finding was that perceptual learning involves improving perceivers' ability to differentiate objects in the environment by attending to features of the stimulus that may not have been previously attended to. Since then, evidence of perceptual learning in the visual domain has been found in a variety of tasks, ranging from simple ones like discriminating line offset (Fiorentini & Berardi, 1980) to more complex ones like object search (Leonards et al., 2002) and facial recognition (Apps & Tsakiris, 2013). Perceptual learning has been documented in the domains of smell and taste, in tasks ranging from discrimination of odors to identifying different brands of beer based on their flavor characteristics (see Hall (2008) for a review). It has also been shown for auditory perception in simple tasks such as discriminating pitch and duration differences as well as more complex ones such as learning contrasts between non-native speech sounds (see Goldstone (1998) for a review of general auditory perceptual learning and Samuel and Kraljic (2009) for the case of perceptual learning for speech).

Specificity, generalization, and the neural correlates of perceptual learning

An important question in this literature concerns the physiological changes that underlie the process of perceptual learning and understanding how they might differ across tasks. Is perceptual learning a relatively low-level process, involving basic sensory modifications, or does it include higher-level, cognitive changes? The literature provides evidence for both types of physiological adaptation. On the one hand, the specificity of perceptual learning

— its failure to transfer — meshes well with the observation that early stages of processing tend to be specific to so-called 'low-level' features like stimulus position, while generalization of learning is consistent with higher levels of processing (Fahle, 2005). Interestingly, as Fahle (2009) observes, the claim that perceptual learning involves modification of low-level sensory cortices would have been quite controversial just three or four decades ago, as it was widely believed that the primary cortices were hard-wired and served simply to extract information from the environment. The reasoning was that if low-level processing were modified during learning, this could negatively impact performance in other, unrelated tasks. However, by the 1980s mounting psychophysical and neurological evidence began to turn the tide against this consensus. For instance, Fiorentini and Berardi (1980) found that subjects were able to improve dramatically when trained to detect line offset with two near-collinear bars (Vernier task), but that performance returned to baseline as soon as the stimulus was rotated by 90 degrees. Gilbert et al. (2009) point out that this kind of specificity provides evidence that early levels of processing must be involved, since we know that early sensory areas tend to be highly specific for basic attributes of stimuli like position in the visual field, orientation, or luminance.

Physiological changes due to perceptual learning may vary depending on aspects of the stimulus and its presentation, suggesting that distinct neural mechanisms may be involved depending on exactly *what* is being learned. In their review, Gilbert et al. (2009) document a number of different types of changes that have been shown to occur in early sensory areas due to perceptual learning. One mechanism involves expansion of cortical representation, thus allowing a larger amount of brain territory to be allocated to the trained area. For example, training in acoustic frequency discrimination can expand the size of the cortical area in the primary auditory cortex that represents the trained frequencies. However, changes in cortical recruitment would predict transfer of learning to related tasks and stimuli, so failure of generalization of perceptual learning is evidence against this mechanism. Alternatively, perceptual learning can also involve changes in the stimulus selectivity of neurons: rather than increasing the number of neurons involved, it means increasing their sensitivity to task-relevant aspects of the signal. Another mechanism involves enhanced neural response due to perceptual learning , either by increasing the activity of individual neurons or increasing the number of recruited neurons. For instance, a study of tone discrimination in owl monkeys has shown an initial increased neural response across all frequencies (not just the trained ones), followed by more differentiated responses to non-target and target stimuli, suggesting that increased neural response may be an early change that precedes sharpening of neural response tuning curves (Li & Gilbert, 2009). Finally, neural plasticity due to perceptual learning may manifest as a change in neural firing patterns, resulting in more synchronous activity across a population of neurons(Li & Gilbert, 2009).

The cases reviewed above primarily yield evidence of low-level, task-specific adaptation in the primary sensory cortices. However, studies have also shown that perceptual learning can involve changes in higher-level processing and can generalize to new tasks and stimuli. For instance, Fahle (2005) highlights several instances involving tasks such as reading, visual object search, or playing action video games, where subjects may learn general strategies that

can transfer to different tasks. Both locus of learning and generalization may be affected by the complexity of the task and the stimuli involved. In visual learning, for instance, simpler tasks may take more time, involve smaller amounts of learning, and be less likely to transfer to novel situations (Fahle, 2005; Fine & Jacobs, 2002). This basic pattern suggests changes in early stages of cortical processing (Fahle, 2005). On the other hand, complex tasks are often learned quickly, show more dramatic improvement, and generalize more readily The distinct patterns of behavior seen in these two cases suggest that simple tasks involve relatively low-level adaptation while more complex tasks may involve higher-level changes. perceptual learning may also involve a combination of both types of adaptation. For instance, Fahle (2009) notes that two distinct patterns of learning have been found for complex visual search tasks: subjects learn both specific stimulus features that do not generalize to other tasks as well as a search strategy that does generalize. In other words, perceptual learning in complex tasks may involve not only homing in on specific features and developing richer object representations but also involves improved sensitivity to context and the kinds of general response strategies required.

1.3 Phonetic learning

Is perceptual learning for speech unique?

Many of the characteristic features of general perceptual learning are also evident in the literature on perceptual learning for speech. For instance, a key feature of perceptual learning across domains is that it is relatively-long lasting. Perceptual learning for speech appears to share this general property. Kraljic and Samuel (2005) found that perceptual learning following exposure to a speaker with an artificial accent persisted 25 minutes after exposure — there was no evidence that learning had attenuated in the intervening time period (in fact, the size of the training effect was numerically larger). A study by Eisner and McQueen (2006) found that perceptual learning for a similar type of artificial accent could persist even longer — they found robust training effects 12 hours after initial exposure, including if participants had slept during this period. Xie, Earle, et al. (2018) found the same effect for adaptation to naturally-accented speech — training benefits persisted 12 hours after exposure. Thus, perceptual learning for speech across different tasks appears to meet the criterion of relative permanence that characterizes perceptual learning in other areas.

Another general feature of perceptual learning that we have already noted is a high level of specificity to the trained stimulus — learning does not tend to generalize beyond the exact exposure context. However, this appears to vary depending on the type of task: simple tasks tend to be highly specific to the training context, whereas more complex ones may generalize. In the case of perceptual learning for speech, results are mixed. There are certainly cases where learning appears to be highly specific. For instance, Eisner and McQueen (2005) show that listeners resist generalization of learning following exposure to a speaker with an ambiguous fricative pronunciation. They found no cross-speaker transfer for this sound

under normal testing conditions. However, subsequent studies have found that generalization can occur for these sounds under particular contexts. For recalibration studies of the type pioneered by Norris et al. (2003), generalization seems robust for certain classes of sounds such as stop consonants (Kraljic & Samuel, 2006, 2007) . It also seems to be sensitive to the degree of acoustic similarity between exposure and test contexts — learning is more likely to transfer when these are highly similar (Kraljic & Samuel, 2005; Reinisch & Holt, 2014). This is consistent with the general perceptual learning literature, where similarity between training and test stimuli is an important predictor of learning transfer.

In the case of speech, generalization may crucially depend on the type of exposure context. For instance, several studies of natural accent accommodation have shown that learning can transfer to new speakers if listeners are exposed to sufficient input variability. Multi-speaker training regimens in such cases have been shown to facilitate speaker-independent learning (Baese-Berk et al., 2013; Bradlow & Bent, 2008). In the general literature on perceptual learning , it has been suggested that the specificity of learning may depend on what is being learned — relatively simple tasks (e.g., such as line offset discrimination) tend to be highly specific, whereas more complex tasks tend to generalize more freely. Moreover, distinct patterns of generalization may be observed for different aspects of complex tasks. For instance, listeners may be able to generalize learning of a global task strategy, but not specific stimulus attributes (Fahle, 2009). Results showing that listeners can achieve accent-independent perceptual learning in some cases (Baese-Berk et al., 2013) suggests that this can hold for speech as well: listeners in these learning conditions appear to acquire a general adaptation strategy that may be independent of the specific phonetic patterns previously encountered. Perceptual learning of accented speech may thus match the pattern of results that have been observed for other complex tasks in non-language contexts.

Finally, research suggests that perceptual learning is relatively automatic and distinct from other forms of learning such as declarative or procedural learning (Fahle, 2001). This appears to hold in some cases of perceptual learning for speech as well. For instance, several studies suggest a dissociation between listeners' top-down expectations or beliefs and whether learning generalizes or not. For instance, Eisner and McQueen (2005) found that learning of a atypical fricative pronunciation generalized when fricatives from the original speaker were cross-spliced into a new voice. Crucially, this occurred despite the fact that the majority of listeners knew that they were perceiving a novel speaker. Analogously, Reinisch and Holt (2014) found that listeners successfully transferred perceptual learning of a similar fricative pronunciation to a new speaker, even though the majority reported this speaker as having a different accent. Even more strikingly, Xie and Myers (2017) found that listeners failed to generalize learning following exposure to Mandarin-accented English to a new speaker, even when they believed that they were hearing the exact same speaker in exposure and test contexts. Finally, Witteman et al. (2015) showed that listeners showed rapid short-term adaptation for a Hebrew-accented Dutch even when the exposure period required them to be engaged in an unrelated task. Overall, such results indicate the perceptual learning for speech happens automatically and in some cases may even be independent of listeners' explicit beliefs about the atypical speech they encounter.

Key questions in perceptual learning for speech

In their review, Samuel and Kraljic (2009) define perceptual learning for speech as "a change in subsequent language processing" following exposure to "speech that is in some way non-canonical, or different than usually experienced." (p. 1208). However, they distinguish between two bodies of literature that approach the phenomenon from somewhat different perspectives. So-called "Theme I" studies that fit classical definitions of perceptual learning, where exposure to a stimulus leads to improved ability to detect or discriminate that stimulus. These include learning of non-native phonetic contrasts, foreign accents or unfamiliar dialects, and degraded (e.g., noisy) speech. "Theme II" studies consist of a recent body of literature around phonetic recalibration, where exposure to atypical speech sounds alters listeners' perception of subsequent sounds but does not necessarily yield improvements in discriminating them. Typically such studies involve presenting listeners with a phonetically ambiguous sound embedded within a disambiguating word frame, with learning shown by a subsequent shift in phonetic categorization.

One benefit of these latter types of studies, the authors note, is that they clearly indicate what is changing in language processing, whereas studies that use comprehension as a measure of perceptual learning don't provide as much insight into what is going on 'under the hood'. However, one potential drawback of phonetic recalibration studies is that the mechanisms they illustrate may not actually be the same as those underlying the improved comprehension and/or processing of accented or atypical speech. Although this has been a common assumption in the literature on perceptual learning for speech, there is actually little evidence to support a direct link between category structure and improved accent perception. Increasingly, studies have begun to question this assumption (Charoy, 2021; Reinisch & Holt, 2014; Xie et al., 2017; Zheng & Samuel, 2020). However, results remain inconclusive, in part due to the different methodologies used across studies. A major goal of this dissertation is to build on such prior work in perceptual learning of accented speech. Specifically, the focus of this project is to investigate the mechanisms that constrain exposure-induced changes to phonetic category boundaries, and to test whether these same mechanisms affect processing of accented speech and the generalization of learning to new speakers.

Perceptual learning of unfamiliar accents and dialects

Much of the literature on perceptual learning for speech has focused on the problem of how listeners adapt to atypical speech in the form of a non-native accent or unfamiliar regional dialect. Accented speech is often initially difficult for listeners and incurs a processing cost (Munro & Derwing, 1995) but with exposure this difficulty rapidly diminishes as listeners adapt to the pronunciation (Clarke & Garrett, 2004). Many studies have found mprovements in accent comprehension and/or processing after relatively minimal minimal exposure to accented speech (Baese-Berk et al., 2013; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Melguy & Johnson, 2021; Vaughn, 2019; Weil, 2001; Xie, Weatherholtz, et al., 2018). These studies on perceptual learning of accented speech suggest that sufficient exposure allows

listeners to adapt to a speaker or group of speakers. However, a drawback of this body of research is that it does not offer a clear answer regarding the mechanisms that underpin perceptual learning for accented speech—learning is assessed simply by measuring whether comprehension improves or not. This metric doesn't give us much direct insight into questions about what is learned and the mechanisms involved in generalization—these must be inferred. Moreover, because these studies use speech from actual accented speakers, they are not able to control precisely for particular phonetic features. These must be inferred from the L1 backgrounds of the speakers, a tricky problem given how variable accented speech has been shown to be (Wade et al., 2007).

Phonetic recalibration

A promising approach to phonetic learning avoids this problem by utilizing artificial accents to address questions on the mechanisms involved in perceptual adaptation to accent. Such studies involve exposing listeners to a phonetically ambiguous pronunciation of a sound that is disambiguated by the context in which it occurs, and perceptual learning is indicated by a subsequent categorization boundary shift on a phonetic continuum. For instance, this might involve replacing the final sound in a word like *moss* with a sound between [s] and [f], where the lexical context leads listeners to interpret the sound as [s] (because *moss* is a real word and *moff* is not). With sufficient exposure to such examples, listeners will shift their categorization boundary of an [s]-[f] phonetic continuum such that more tokens are categorized as /s/, effectively expanding the size of the phonetic category at the expense of /f/. This type of perceptual learning has been referred to as phonetic recalibration because it involves targeted adjustment of a single phonetic category. Seminal phonetic recalibration studies (Bertelson et al., 2003; Norris et al., 2003) showed that various types of disambiguating contexts could be used to achieve perceptual learning of phonetically ambiguous segments. Norris et al. (2003) used a lexical decision task to expose Dutch listeners to an ambiguous segment [?] that was midway between [s] and [f]. They separated their listeners into groups and presented each with the same sound but in a different word context, such that one group was led to interpret the sound as /f/ while the other group was led to hear it as /s/. They found that after exposure these two groups showed phonetic categorization shifts in opposite directions—those trained to hear [?] as an [f] shifted the boundary toward /s/, while those trained to hear it as [s] shifted the boundary toward [f]. Bertelson et al. (2003) showed that phonetic recalibration could be achieved using a different methodology that used visual information as the disambiguating context. They exposed listeners to auditory [ada], [aba], or [a?a], where the ambiguous sound was midway between [d] and [b]. They paired these audio recordings with videos of a face articulating either [aba] or [ada]. They found that participants perceived the ambiguous sound based on what they saw articulated in the video, leading to category boundary shifts analogous to what was observed in Norris et al. (2003).

Phonetic recalibration vs. accent accommodation

One limitation of the paradigms used in the phonetic recalibration studies discussed above is that they often fail to provide evidence that exposure to a particular accent improves perception of that accent, as measured by improved subsequent processing and/or comprehension scores. While it is often assumed that retuning of phonetic categories may be a mechanism for learning a foreign accent or unfamiliar regional dialect, few studies have tested this question empirically. Reinisch and Holt (2014) made important early steps by demonstrating that phonetic recalibration could occur in the context of globally-accented speech. They found that listeners were able to learn an ambiguous pronunciation of a target fricative sound when it was embedded in a Dutch-accented voice, and that listeners were able to generalize learning to novel speakers with the same accent. Their results suggest that the phonetic recalibration paradigm, which involves targeted manipulations of individual critical phonemes, may be a viable model for investigating perceptual adaptation to accent. However, more recent tests of the question have yielded mixed results. Xie et al. (2017) did find a connection between changes in category structure and facilitated lexical processing, but they used a natural accent as opposed to the typical ambiguous artificial accents used in recalibration studies. Zheng and Samuel (2020) found both recalibration effects (category shifts) and changes in lexical processing of a Mandarin-accented speaker with a similar /s/-like pronunciation of /θ/. However, they found no correlation between category shifts and accent accommodation. Finally, Charoy (2021) investigated whether category shifts could be found for both ambiguous pronunciations (/θ/ = [θ/s]) and substitutions or "bad maps" (/θ/ = [s]). She argued that both types of pronunciations may occur in naturally-accented speech, so if recalibration facilitated accent accommodation it should be possible to observe categorization shifts in both types of scenarios. In an initial set of experiments, the author found that category shifts were indeed observed for both pronunciations. However, follow-up experiments failed to find a clear relationship between such category shifts and lexical processing of critical words (/θ/ = [θ/s] or [s]) produced by the same speaker. Previous accent exposure did not appear to reliably facilitate processing of such accented items.

1.4 The current study

The primary goal of this dissertation is to investigate the mechanisms that underlie perceptual learning of accented speech, with a focus on better understanding the relationship between changes in phonetic category structure, lexical processing, and subsequent perception of novel speakers. In testing these questions, this project hopes to bring together two bodies of literature that we have reviewed on perceptual learning for speech: studies on perception of natural accents, which have generally used comprehension- or processing-based measures to evaluate perceptual learning, and phonetic recalibration studies utilizing artificial accents, which measure learning via categorization shifts following exposure to an ambiguous pronunciation of a target sound. If there is indeed a relationship between cate-

gory shifts and accent processing, then we ought to observe improved word recognition for critical words containing the trained accent.

Each of the following chapters investigates an aspect of perceptual learning for accented speech, focusing on the mechanisms underlying changes to phonetic category structure, lexical processing, and the generalization of learning to novel contexts. Together, this set of studies aims to contribute to the broader question of whether recalibration of specific phonetic categories may underlie accent accommodation. Chapter 2 investigates the mechanisms behind phonetic recalibration. Existing literature provides support for two possible mechanisms by which listeners recalibrate phonetic categories following exposure to an atypical pronunciation — listeners can either shift the target category or expand it in perceptual space. Crucially, both mechanisms can account for the recalibation effect, but existing literature does not providing conclusive evidence in favor of one or the other. Chapter 3 focuses on the connection between changes to category representations and online lexical processing, directly testing the common assumption that the former facilitates the latter. Chapter 4 examines how phonetic learning may generalize to novel speakers. These experiments test whether the same mechanisms that underlie phonetic learning of a single speaker also generalize to novel speakers.

Chapter 2

Mechanisms of perceptual adaptation to a novel accent

2.1 Introduction

As listeners, we sometimes encounter speech that deviates from what we typically hear in our everyday lives.[1] Whether due to a non-native accent or regional dialect, or an idiosyncratic pronunciation, such variation poses a potential challenge for listeners, resulting in decreased comprehension and/or increased processing time. Fortunately, listeners can rapidly adapt to an atypical pronunciation, reducing or eliminating this 'accent cost' with sufficient exposure to the speaker (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Melguy & Johnson, 2021; Sidaras et al., 2009; Vaughn, 2019).

Such perceptual adaptation is well-established in the literature, and has been found using a variety of tasks. Clarke and Garrett (2004), for instance, exposed listeners to Spanish- or Chinese-accented English where the final word in the sentence was unpredictable, and listeners' task was to indicate if a visual word target on their computer screen matched the spoken word or not. By the end of the task, subjects' responses matched the baseline found for natively-accented speech. Another study by Bradlow and Bent (2008) measured sentence transcription accuracy of Chinese-accented speech for trained and untrained groups of listeners, finding that listeners previously exposed to the accent showed improved accuracy vs. controls, with listeners trained on multiple speakers able to generalize learning to a new speaker of the same accent. Sidaras et al. (2009) utilized a similar paradigm, exposing

[1]This chapter was previously published as: Melguy, Y.V and Johnson, K. (2022). Perceptual adaptation to a novel accent: Phonetic category expansion or category shift? *Journal of the Acoustical Society of America*, 152(4), 2090-2104. https://doi.org/10.1121/10.0014602. The contribution of each author is as follows: Yevgeniy Melguy developed the research question, produced the experimental materials, collected data, and drafted the manuscript; Keith Johnson produced Figure 2.2 and contributed a description of acoustic and perceptual similarity for the critical fricative sounds in this study, in addition to assisting with various aspects of experimental design, statistical analysis, and experimental implementation. We thank Arthur Samuel and an anonymous reviewer for their helpful feedback on a draft of this manuscript.

listeners to sentences produced by 6 speakers of Spanish-accented English, and then testing transcription accuracy on novel sentences and words produced by either 6 different speakers or the same group of speakers they heard in training. Trained listeners showed improved performance for both the familiar and novel speaker group vs. untrained controls. Melguy and Johnson (2021) also found evidence of adaptation to a speaker of Mandarin-accented English, as measured by improved sentence transcription accuracy over the course of a 60-trial exposure period. Vaughn (2019) obtained a similar result for Spanish-accented English, finding a significant improvement in transcription accuracy across just 40 sentence-level trials.

The common pattern in such studies is that listeners can rapidly adapt to a specific speaker's accent following a relatively brief period of exposure, a result that is consistent across different measures of language processing, from word or sentence transcription accuracy (Alexander & Nygaard, 2019; Bradlow & Bent, 2008; Melguy & Johnson, 2021; Sidaras et al., 2009; Vaughn, 2019) to reaction times in online lexical processing tasks (Clarke & Garrett, 2004; Xie & Myers, 2017; Xie et al., 2017).

Phonetic recalibration

How do listeners achieve such adaptation? One commonly assumed mechanism is lexically-guided recalibration or 'retuning' of phonetic category boundaries, which allows the perceptual system to dynamically adjust to changes in the linguistic environment (Norris et al., 2003). Such adaptation facilitates subsequent speech processing by adjusting prelexical categories to more closely match recently encountered tokens, thereby improving accuracy for a particular speaker or accent. In their seminal study, Norris et al. (2003) argued that lexical feedback could allow listeners to perceptually adapt to a novel regional dialect or accent via adjustment of phonetic categories. They created an artificial accent by mixing together two sounds [s] and [f] and embedding the resulting ambiguous sound [s/f] in naturally recorded words, replacing either /s/ or /f/. Following exposure to this pronunciation, listeners shifted their categorization boundary on an [s]-[f] phonetic continuum, with the direction of the shift depending on the training condition (whether they heard [s/f] in /s/ or /f/ words). Subsequent studies have added several key findings about the nature of recalibration. First, they show that phonetic recalibration is often speaker-specific, failing to transfer to a new speaker with the same pronunciation (Eisner & McQueen, 2005; Kraljic & Samuel, 2005). Authors of these studies have argued that acoustic similarity between training and test pronunciations constrains the likelihood of generalization, as the phonetic realization of fricatives can differ substantially between speakers. This observation sees further support in later recalibration research involving voiceless fricatives, which has shown that cross-speaker generalization is possible when exposure and generalization speakers' productions span a similar range of perceptual space (Reinisch & Holt, 2014). In other cases generalization appears to occur more freely, both across speakers and across categories for the same speaker. For instance, Kraljic and Samuel (2006) showed that learning for an ambiguous pronunciation /d/ = [d/t] transferred to a novel speaker and to a novel place of articulation – trained listeners classified

more sounds as voiced on both [d]–[t] and [b]-[p] phonetic continua. The authors also explain this generalization asymmetry between stops and fricatives in acoustic terms - the phonetic implementation of voicing differs minimally across speakers and across places of articulation, whereas fricatives show much larger differences and thus tend to resist generalization.

Schuhmann (2014) found that within-speaker generalization was possible with fricatives in some cases: listeners trained on an ambiguous voiceless fricative pronunciation [f/s] generalized the same pattern to a voiced fricative contrast [v]-[z]. She argued that generalization occurred, despite the voicing difference, because both [f]-[s] and [v]-[z] contrasts involve similar acoustic-phonetic cues to place of articulation. Generalization was also found by Mitterer et al. (2016), who trained listeners on a pronunciation involving tensified (underlyingly lax) Korean stops that were phonetically ambiguous in place of articulation (e.g., /t/=[t*/p*]). They found that listeners generalized learning to phonetically similar non-tensified lax stops [t]-[p], but not to the more distinct aspirated stops [t^h]-[p^h], concluding that acoustic similarity was an important predictor for generalization of learning.

Additional research suggests that listeners can also generalize learning across prosodic positions, but again such transfer seems to be tightly constrained. For instance, Jesse and McQueen (2011) found that listeners trained on an ambiguous fricative pronunciation [f/s] in word-final position generalized learning to word-initial instances of the sound, but Mitterer et al. (2013) failed to find similar transfer of learning for Dutch liquids which, unlike fricatives such as /s/ or /f/, are realized differently in onset vs. offset positions. They trained listeners on an ambiguous liquid pronunciation involving word-final /r/ and /l/, where /r/ can occur as an approximant, while /l/ is velarized (e.g., /l/ = [ɹ/ɬ]). They then tested generalization of learning to different allophonic variants of these sounds in word-final and word-initial positions. They found no generalization of learning in either case, concluding that learning occurred at the level of position-dependent allophones and that generalization is constrained by acoustic similarity with exposure sounds. A more recent study by Mitterer and Reinisch (2017) also failed to find generalization of learning across word position for German stops. Listeners were exposed to (underlyingly) voiced stops that were ambiguous in place of articulation. These sounds occurred word-finally and, due to a German devoicing process, were realized as phonetically voiceless (e.g., /d/ = [t/k]). Results showed that listeners generalized learning to phonetic contrasts involving (underlyingly) voiced stops [d]-[g] as well as (underlyingly) voiceless stops [t]-[k] in word-final position. However, there was no generalization when the same sounds occurred word-medially. Again, the conclusion was learning is contextually-specific and dependent on acoustic similarity between exposure and test segments. Reinisch and Holt (2014) also found that cross-speaker transfer of perceptual learning for ambiguous fricative pronunciations was possible, provided that exposure and generalization speakers' productions spanned the same range of perceptual space. This mirrors recent findings from natural accent accommodation studies (Alexander & Nygaard, 2019; Xie & Myers, 2017), which suggest that acoustic similarity between exposure and test speakers is a key ingredient for successful generalization of perceptual learning to novel speakers.

The current body of work on phonetic recalibration has shown that the perceptual system

is flexible and able to adapt to atypical pronunciations of a wide variety of speech segments, as well as to generalize such learning to novel segments. However, generalization of learning appears to be tightly constrained by phonetic similarity and the distribution of relevant acoustic cues, which may differ across contexts, varying by word position, by segment, and by speaker. This mirrors recent findings from natural accent accommodation studies Alexander and Nygaard, 2019; Xie and Myers, 2017, which suggest that acoustic similarity between exposure and test speakers is a key ingredient for successful generalization of perceptual learning to novel speakers. It also shows that such learning occurs at the prelexical level and can generalize to novel segments, prosodic positions, words, and speakers.Moreover, recalibration has been shown to yield changes not just in categorization behavior, but also in online measures of lexical processing and listener judgements of category goodness (Eisner et al., 2013; Xie et al., 2017).

Mechanisms of phonetic learning for accented speech

Despite extensive evidence that listeners can successfully 'retune' or 'recalibrate' perceptual categories following exposure to an atypical pronunciation (Kraljic & Samuel, 2005, 2006; Mitterer et al., 2013; Norris et al., 2003; Reinisch et al., 2013), and that perceptual categories have internal structure (Iverson & Kuhl, 2000; J. L. Miller & Volaitis, 1989), there is no consensus about the underlying mechanisms involved. In this study, we explore two possible adaptation strategies that have been proposed in prior research: phonetic category boundary shift vs. category boundary expansion (e.g., (Kleinschmidt & Jaeger, 2015)). Under the first account, listeners are said to make targeted adjustments to phonetic category representations following exposure to a non-canonical pronunciation. Such shifts are expected to be specific to the sound pattern involved in training with no expected generalization of learning to other sound categories. Under the second, listeners may simply become more tolerant of atypical pronunciations of a given category. This process has been described as a general relaxation of default phonetic categorization criteria. This means that listeners may generalize learning beyond the specific phonetic pattern in the exposure accent, accepting multiple different realizations of the trained category as instances of that category. Importantly, such expansion could be uniform or non-uniform. A uniform expansion of the trained category would involve a general broadening of the category in phonetic space, potentially altering subsequent perception of all neighboring sounds. A non-uniform expansion, by contrast, would involve expansion of just one part of the category toward the region of phonetic space where listeners had encountered the unusual tokens of the target.[2]

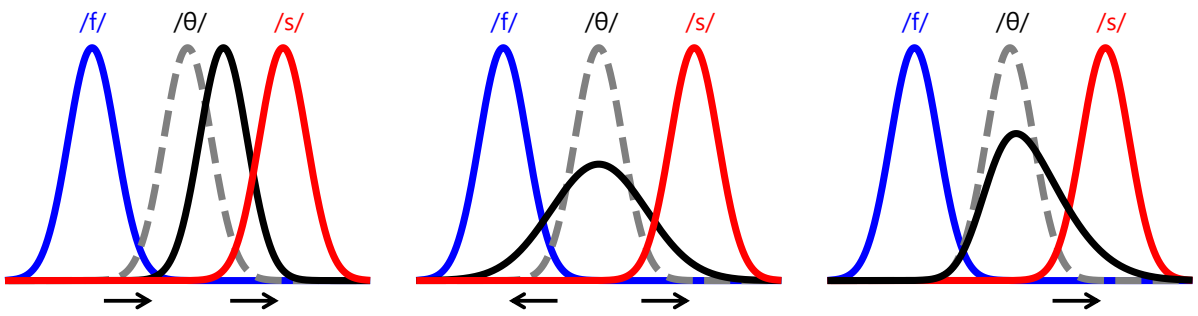
[2]These different views on how recalibration might occur see a parallel in earlier research on selective adaptation (SA), a perceptual phenomenon where the repeated presentation of a given stimulus (e.g., /ga/) reduces subsequent perception of that stimulus (resulting in less /ga/ responses when listeners categorize a [ga]-[ka] phonetic continuum). The earliest account of SA Eimas and Corbit, 1973 posited that that the basic mechanism was fatigue of phonetic feature detectors. So, for instance, a detector tuned to the feature [+voice] was predicted to show a reduction in output strength following presentation of a sound like [ga], across the *entire* range of VOT values to which that detector is sensitive. This view thus suggests a

These different mechanisms are schematized in Figure 2.1, where we illustrate three possible ways by which listeners may restructure their /θ/ category boundaries following exposure to an artificial accent where /θ/ = [θ/s]. Here we show hypothetical perceptual category distributions for /θ/ and two neighboring fricatives, /f/ and /s/, which are similar to it perceptually (Cutler et al., 2004; G. A. Miller & Nicely, 1955) and acoustically (Stevens, 1998). Crucially, we assume that /θ/ lies in between these two categories in perceptual space (see Section 2.1 for a detailed discussion of these sounds and a justification for placing them along a continuum of perceptual similarity). If listeners recalibrate by *shifting* the category (change in mean, see panel A), we expect a shift in both sides of the /θ/ distribution toward /s/. If listeners utilize a *uniform expansion* strategy (increase category variance, see panel B), then we expect a shift on both sides of the category distribution, toward /s/ on the right and toward /f/ on the left. Finally, if recalibration occurs via *non-uniform expansion* (added skew of the distribution, see panel C), then we expect to see a shift in the category boundary toward /s/ but no change in the /f/-/θ/ boundary.

Previous work on phonetic recalibration has generally assumed a category shift mechanism – a direction-specific adjustment of the category boundary between the two sounds involved in creating the artificial accent. However, the general method of testing for a perceptual learning effect in such studies usually does not allow us to distinguish between a shift of one category toward another versus a general broadening of that category in phonetic space.

Previous work has noted the possibility that listeners could be relying on either category shift or expansion to achieve recalibration. As Kleinschmidt and Jaeger (2015) point out, both are plausible mechanisms that can account for the recalibration effect: listeners can either shift a category in phonetic space (change the mean) or expand it (increase the variance). Moreover, they argue that both strategies are available to listeners, and both may be useful depending on the speech context and listeners' prior assumptions about how pronunciation varies across sound categories and across speakers – note, however, that their modeling assumes that categories have a normal distribution defined by just two parameters (mean and variance), and they do not discuss skewed distributions. Yet despite the extensive body of work on phonetic recalibration, evidence to support a category category shift versus some version of category expansion is lacking – existing findings are compatible with all three possible strategies proposed here. This is because most recalibration studies have either focused on testing just the sound pattern listeners heard during exposure (Eisner & McQueen, 2006; Norris et al., 2003), testing generalization of the same pattern to a different speaker (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Reinisch & Holt, 2014), or looking for cross-series generalization of the trained sound pattern to another pair of sounds related on some feature dimension (Kraljic & Samuel, 2006; Mitterer et al., 2016; Mitterer & Reinisch, 2017; Reinisch & Mitterer, 2016; Schuhmann, 2016). However, despite this ambiguity in the

more general mechanism structurally parallel to uniform category expansion. An alternative to the feature detector explanation was the adaptation-level (AL) or contrast account of SA (Diehl, 1989; Diehl et al., 1978). Under this view, repeated presentation of a given stimulus (the anchor, or adaptation level) causes listeners to shift their categorization boundary toward it, a mechanism more closely resembling the category shift or the non-uniform category expansion hypotheses being considered in the present study.



(a) (b) (c)

Figure 2.1: Possible recalibration strategies following exposure to an ambiguous /θ/ = [θ/s] pronunciation (dotted line indicates category distributions prior to accent exposure): (a) shows recalibration by category shift, while (b) and (c) show recalibration by uniform or non-uniform category expansion. Category shift (a) and uniform expansion (b) predict that both the /f/-/θ/ and the /θ/-/s/ boundaries will shift after exposure, while non-uniform category expansion (c) allows for only a shift on the /θ/-/s/ boundary.

existing recalibration literature, other research on perceptual learning more generally does offer some evidence for each of these strategies.

Category Shift

For instance, several studies utilizing either natural or artificial accents have provided evidence for the type of pattern-specific learning that is suggested by a category shift mechanism. For instance, Maye et al. (2008) trained listeners on an artificial accent where front vowels were systematically lowered (e.g., *witch* → *wetch*). Following exposure, listeners were more likely to process novel words with lowered front vowels (matching the exposure pattern) as real words. However, listeners who were tested on a novel accent where front vowels were raised did not show increased acceptance of these items as real words, suggesting pattern-specific learning. Analogously, Bradlow and Bent (2008) showed that exposure to multiple talkers of Chinese-accented English led to improved transcription accuracy with a novel Chinese-accented speaker, but not a Slovakian-accented one. This suggests that listeners learned an accent-wide schema that was specific to the common phonetic characteristics of Chinese-accented English, but not Slovakian-accented English. Alexander and Nygaard (2019) also found relative specificity of learning, which they took as evidence for a boundary shift learning mechanism — listeners who were trained on single-word utterances produced within a given accent (either Korean- or Spanish-accented English) and subsequently tested on a novel speaker of the same accent saw a significant training benefit, while those tested on a novel speaker of a different accent did not. Meanwhile, results from multiple-accent training were mixed and yielded a smaller training benefit that was less consistent, showing

up for Spanish-accented tokens but not Korean-accented ones. Subsequent analyses of vowel productions across the different accents suggested that acoustic similarity between exposure and test items may have facilitated generalization where this was observed, regardless of training/test accent pairings.

Category Expansion

Other studies show evidence for a less specific learning strategy that may support a version of the category expansion mechanism proposed here. This kind of general strategy has been presented as either relaxation of phonetic categorization criteria (e.g., (Zheng & Samuel, 2020)) or as general category expansion (Kleinschmidt & Jaeger, 2015; Schmale et al., 2012). The idea is that rather than shifting phonetic category boundaries in a specific direction, listeners may just expand them, which may help them with diverse or unfamiliar accents. Lev-Ari (2015) presents an analogous view of non-native speech comprehension – listeners expect non-native speakers to be less competent in general, and therefore adjust processing at all levels (phonetic, lexical, syntactic, etc.) to accept more deviations from native-speaker norms. Evidence for this mechanism is shown by generalization of perceptual learning beyond the specific segmental patterns listeners are exposed to. For instance, White and Aslin (2011) found that infants trained on a vowel fronting pattern /ɑ/ → /æ/ (*dog* → *dag*) showed some generalization of learning to a similar, but distinct pattern of fronting /ɑ/ → /ɛ/ (*dog* → *deg*). The authors suggested that accent exposure may have resulted in category expansion, but that this was constrained by similarity to the trained pronunciation. Weatherholtz (2015) found evidence for more general category expansion, showing that listeners exposed to a novel pattern of back vowel lowering generalized learning to a novel accent where back vowels were raised (contrary to the findings of (Maye et al., 2008)), as well as to a structurally parallel pattern where front vowels were lowered. However, listeners who were exposed to a pattern of back vowel raising showed no generalization to raised front vowels – learning was specific to the exposure pattern. Such findings suggest that while phonemes may expand in phonetic space due to accent exposure, there are still clear constraints on the degree to which listeners are willing to relax categorization criteria and generalize beyond the specific phonetic pattern(s) they are exposed to.

The current study

The primary aim of this study is to investigate the mechanisms responsible for the change in a phonetic category boundary observed in a typical lexically-guided recalibration experiment. Crucially, we assume that in either scenario listeners are making an adjustment to a phonetic category representation following exposure to an atypical realization of that category. The question is whether such recalibration is contrast-specific (boundary shift) or more general (category expansion).

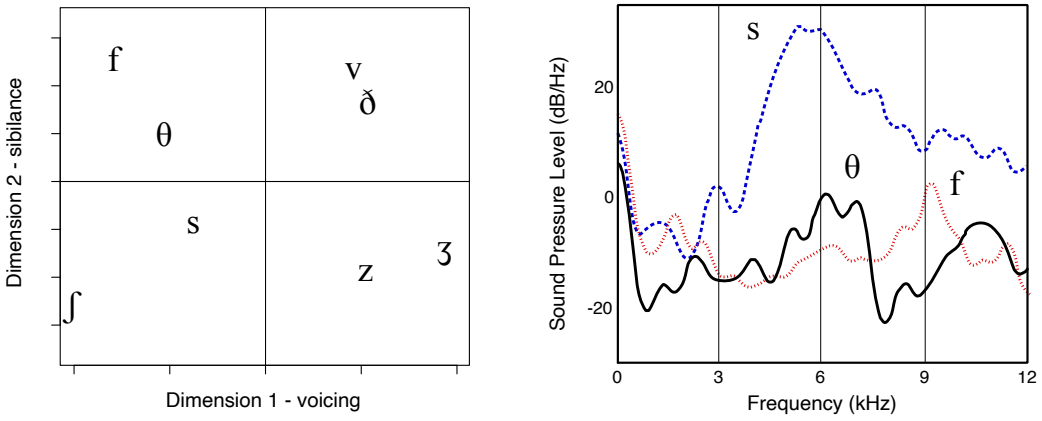


Figure 2.2: Perceptual space (left panel) and fricative noise spectra (right panel) illustrating the similarity of the non-sibilants [θ] and [f] and of the coronal fricatives [s] and [θ]. [θ] is plotted with a solid black line, [s] is plotted with a dashed blue line, and [f] is plotted with a red dotted line.

Fricatives in perceptual and acoustic space

The specific sounds used in training and test phases of the current study were selected based on their close correspondence to actual pronunciation differences in non-native speakers of English, as well as their proximity to one another in perceptual and acoustic space. Because the dental fricative /θ/ is a cross-linguistically unusual sound typically absent from language phoneme inventories, L2 learners across different language backgrounds struggle to realize it in a native-like manner. Even within a group of L2 speakers sharing a common L1 background, it is normal to see a wide range of variation in how this segment gets realized. Two frequent substitutions involve other (more typologically common) fricatives that are phonetically close to [θ], such as [s] or [f] (Seibert, 2011).

The experiments reported here thus make use of the fricatives [f], [θ], [s] (experiment 1), and [ʃ] (experiment 2), and critically assume that [θ] is perceptually intermediate between [f] and [s]. This assumption is justified by perceptual data (see Figure 2.2, left panel). The perceptual space illustrated there was derived from the 0dB SNR confusion matrix published by G. A. Miller and Nicely (1955), using the multidimensional scaling method for confusion matrix data developed by Shepard (1972) (see also Johnson (2003) for a short explication of the method). The first perceptual dimension that emerges in the analysis of English fricative confusions separates the voiced fricatives from the voiceless fricatives. The second

perceptual dimension separates the sibilant fricatives [s, ʃ, z, ʒ] from the non-sibilants [f, θ, v, ð]. Importantly, in the perceptual space [θ] is between [f] and [s]. One key acoustic cue that is related to this perceptual configuration is illustrated by the spectra shown in the right panel of Figure 2.2, which shows acoustic power spectra taken from the midpoint of each of the fricatives [f], [s], and [θ] that were used in experiment 1 below. The sibilance of [s] is apparent in how much louder (vertically displaced) the [s] spectrum is; [f] and [θ] are similar to each other in having lower amplitude. In addition, [s] and [θ] are similar to each other in having a peak amplitude between 5 and 7 kHz. This reflects the resonant frequency of a short resonant tube in front of the tongue constriction of [θ] and [s], which is shorter in [θ]. The effective length of the acoustic cavity in front of the constriction is much shorter in [f], resonating above 9 kHz (Stevens, 1998). The spectrum of [ʃ] (not shown in Figure 2.2 for the sake of clarity) has high amplitude like [s] and lower frequency resonant peaks at about 3 and 5 kHz.

2.2 Experiment 1

The proximity of these two sounds in the phonetic space surrounding [θ] makes for a clear test of a category shift vs. expansion strategy, with potential real-world implications for adaptation to natural accents. We investigated this question by exposing two groups of listeners to the same artificial accent: a pronunciation of the voiceless dental fricative /θ/ that was ambiguous between [s] and [θ]). Previous recalibration literature has used a similar training accent involving /θ/, intended to approximate a common pronunciation of this sound by L1 Mandarin speakers of English (Charoy, 2021; Zheng & Samuel, 2020). Listeners were then tested on different minimal pair continua: one group heard tokens along [θ] – [s] continua (e.g., *think* – *sink*), while the other group heard [θ] – [f] continua (e.g., *thought* – *fought*). Predictions for the changes in categorization under each type of possible recalibration mechanism are summarized in Figure 2.1. If recalibration is achieved by category shift (shift in mean, no change in variance), then we should expect to see a shift toward [s] in the [θ]-[s] test group and a shift toward [θ] in the [θ]-[f] test group. However, if training induces a uniform broadening of the /θ/ category, then we can expect an increase in the proportion of /θ/ responses in both test groups, reflecting listeners' willingness to generally accept more atypical tokens as exemplars of /θ/, regardless of phonetic match to the exposure pronunciation. Finally, if listeners expand the /θ/ category in a non-uniform way following accent exposure (skewed distribution), we would expect to see a shift in the [θ]-[s] group toward [s] but no change in the categorization results for the [θ]-[f] group.

Experimental platform

This experiment was implemented via a custom web program[3], and data was collected via Amazon's Mechanical Turk (Mturk), a platform where workers complete HITs (Human Intelligence Tasks) for compensation. In recent years Mturk has been an increasingly popular platform for conducting linguistics research, including auditory perception tasks (Cooper & Bradlow, 2018; Melguy & Johnson, 2021; Vaughn, 2019; Xie & Myers, 2017) as well as phonetic recalibration tasks (Charoy, 2021; Liu & Jaeger, 2019). In contrast to data obtained from the lab, Mturk offers the benefit of a more demographically diverse participant pool and efficient collection of large amounts of data. Recent studies have shown that Mturk can yield high-quality psychometric data comparable to that obtained from laboratory studies (Buhrmester et al., 2011) despite the inability for researchers to precisely control for factors such as the experimental environment or hardware used for stimulus presentation. Recent research has shown that when differences in audio quality are accounted for, online participants can closely match the quantitative and qualitative patterns of their lab counterparts (Cooke & García Lecumberri, 2021). [4]

Participants

189 participants were initially recruited via Amazon's Mechanical Turk to participate in the main experiment. Exclusion of participants who failed to meet experimental criteria left a total of 120 participants whose data were retained for analysis (see Section 2.2 for an explication of exclusion criteria). All participants lived in the U.S., were native speakers of English, reported normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/hour for their participation. An additional 22 participants were recruited under the same selection criteria for a norming task, which was used to determine the most ambiguous critical /θ/-words to be used in the exposure task. Experimental participants were primarily male (M = 80, F = 37), with 3 subjects declining to report and no subjects selecting the 'other' option. Most participants reported white ethnicity (N=94), with 7 self-reporting as Asian, 9 as black, 6 as Hispanic, and 4 declining to answer. Participants spanned a wide range of age groups (6 = 21-25, 19 = 26-30, 33 = 31-35, 17 = 36-40, 11 = 41 – 45, 10 = 46 – 50, 12 = 51-55, 8 = 56-60, 4 = 61+). Participants self-reported moderate experience with accented English on a 7-point scale (M = 4.64, SD = 1.71, range = 1-7).

[3] A version of this program can be accessed at https://github.com/keithjohnson-berkeley/perception_on_the_web.

[4] As noted by a reviewer, the importance of audio quality may be especially important in the perception of certain sounds such as voiceless fricatives where differences between sounds are primarily in the higher frequencies of the spectrum. While we utilized a headphone check in the present study, it was not possible to precisely control for differences in audio hardware quality, although results suggest that the majority (over 75%) of participants utilized high-quality headphones, based on participant self-reports. Nonetheless, previous research Charoy, 2021; Liu and Jaeger, 2019 has shown that it is possible to conduct successful online recalibration experiments with voiceless fricatives, even when a headphone check is not utilized.

Stimuli

The materials for the exposure task consisted of 100 English words and 100 nonwords. Stimuli were based on the lists used in the lexical decision exposure task from Zheng and Samuel (2020) and were modified to be appropriate for testing the present question (e.g., words containing /f/ were removed from the critical /θ/-items, and additional stimuli were added to bring the total number of items to 200.

All stimuli were recorded in a quiet room at a sampling rate of 44,100 kHz using an Audiotechnica AT2020USB+ condenser microphone. The speaker (27, M) was a native speaker of American English and was living in Berkeley, California at the time. Stimuli consisted of 20 critical words containing an ambiguous /θ/ (phonetically between [θ] and [s]), 20 unambiguous words containing /s/, 60 filler words, and 100 filler non-words (see Appendix A for a full list of training materials). Tokens from the 3 word-lists (/θ/-words, /s/-words, and word fillers) were equated in mean lexical frequency, with an average word frequency in each list of 4.1 occurrences per million in the SUBTLEX-US corpus (Brysbaert et al., 2012). Words ranged in length from 2-4 syllables, and the lists were closely matched in mean syllable length (/θ/-words = 3.1, /s/-words = 3.1, fillers = 3.2). The majority of nonwords (80 items) overlapped with those used in Zheng and Samuel (2020), and were supplemented with 20 additional items, created from filler words by replacing approximately 1 sound per syllable, following the method used in Kraljic et al. (2008). Non-word fillers were also between 2-4 syllables long and matched the real-word items in mean syllable length (3.1). None of the filler items contained the sounds /θ/, /f/, or /s/.

Ambiguous realizations of /θ/ for the 20 critical exposure items were created using STRAIGHT (Kawahara & Morise, 2011), following recent phonetic recalibration work (Reinisch & Holt, 2014; Reinisch et al., 2013). Using STRAIGHT allows morphing of word stimuli in their entirety (rather than excising and mixing just the fricative portions). This means that the resulting morphed stimuli are ambiguous on all cues to consonant identity (spectral characteristics, duration, formant transitions, etc.).

For each critical item, two recordings were made: one naturally-produced /θ/ word (e.g., *empathy*) and one non-word counterpart where /θ/ was replaced with /s/ (e.g., *empassy*). The two recordings were then mixed in their entirety, generating a 7-step phonetic continuum for each item (e.g., *empathy – empassy*). Prior to mixing, temporal anchors were placed at major acoustic landmarks (e.g., the onset of voicing for vowels, the onset/offset of silence for stops) to ensure that segments of the same type were mixed together, and to obtain more natural-sounding morphed stimuli (Reinisch et al., 2013). A norming task (see section 2.2 below) was then used to select the most phonetically ambiguous step along the resulting continuum.

The same morphing procedure was used to create the minimal pairs used in the test phase, following Reinisch and Holt (2014) and Reinisch et al. (2013), who report that using minimal pairs rather than dummy syllables during the test phase reduced between-subject variability and facilitated cross-talker generalization. We created 4 /θ/-/s/ minimal-pair continua and 4 /θ/-/f/ minimal-pair continua. In each case, 2 of the continua had the

critical sound occur word-finally, and 2 had it word-initially (see Appendix A for a full list of Exp.1 test materials). The same speaker used to record training materials also produced all test materials. Finally, both test and exposure items were normalized in amplitude, resampled to 16KhZ, and converted to mp3 format to ensure between-browser compatibility during audio presentation.

Procedure

Pretest

In order to select the most ambiguous critical /θ/ items for the exposure task, the 20 continua were subjected to a pretest by a separate listener group. For this purpose, 22 participants were initially recruited via Mturk. Data from 2 participants was removed due to failure to complete the task, and 1 additional participant was excluded due to inability or unwillingness to perceive the difference between items (i.e., giving the same response for all tokens within a continuum, for over 50% of continua).

For each continuum, the middle 5 steps were selected (the endpoints were not used). Stimulus presentation was blocked by continuum, and 4 lists were created, each with a different random order of continua and step numbers with a given continuum. Participants were randomly assigned to one of these lists. Each participant was thus asked to judge a total of 100 tokens (20 continua x 5 tokens each). Once the participant had read the task instructions, they clicked a button on their web browser to begin the task. Listeners were asked to complete the task in one sitting and to make their response as quickly and accurately as possible. For each trial, listeners heard a token and were asked to indicate whether they heard a real-word (e.g., *empathy*) or a non-word where the /θ/ had been replaced with /s/ (e.g., *empassy*). Responses were made by pressing one of two keyboard keys ('z' = non-word, 'm' = real word). The response labels were present on the screen for the duration of whole trial. Once the listener submitted their response, playback of the next item began automatically (ISI = 1000 ms). The selected ambiguous step number for each continuum was based on the step closest to the category boundary between response options (i.e., the point at which 50% of participants indicated hearing the /θ/-word). In cases where the percentage of /θ/-word responses for this token fell below 50%, the previous step on the continuum was selected (e.g., if token 4 had an average of 45% /θ/-word responses and token 3 had an average of 60% /θ/-word responses, token 3 was selected). This was done to compensate for the real-word bias in speech perception (Ganong, 1980) and to ensure that the target phoneme was phonetically ambiguous, following Reinisch et al. (2013). Results for the selected step closely matched those obtained in their study: the average step number of the ambiguous token was near the middle of the 7-step continuum (3.90), and the mean percentage of /θ/-word responses at this step was 73%.

Screening

Prior to proceeding to the experimental task, all participants completed a screening task designed to ensure that they were wearing headphones. Previous web-based studies have found a significant difference in transcription accuracy based on the quality of audio hardware participants used (Cooke & García Lecumberri, 2021; Melguy & Johnson, 2021). While it is not possible to completely control for the quality of audio hardware with a crowd-sourced participant pool, a headphone check ought to significantly reduce the amount of between-listener variation in audio quality.

For the audio check, we used a custom JavaScript program created by Woods et al. (2017), which utilizes a three-alternative forced choice (3AFC) task with 200 Hz pure tones. In each trial, a random one of the 3 tones is in antiphase across the stereo channels, making it sound significantly quieter when played through headphones, but not when played through loudspeakers. Listeners must decide which of the 3 tones is quietest. This method has been previously demonstrated to be highly effective in both lab and web-based studies with a very small number of trials (see Woods et al., 2017 for more detailed information on stimuli, protocol, and validation results). Questionnaire results suggest that the task was effective, with the vast majority of participants self-reporting use of high-quality headphones (N = 93) or earbuds(N = 24), although as in Woods et al. (2017) a small number purportedly passed the task using loudspeakers (N = 3).

Training

Subjects were randomly assigned to one of two conditions (training vs. no training, cf. (Kraljic et al., 2008)). All participants in the training condition completed a lexical decision task designed to expose them to the speaker's accent. Each heard the same 100 nonwords, 60 filler words, 20 ambiguous /θ/-words, and 20 unambiguous /s/ words. Exposure items were presented in a separate random order for each participant.

Listeners in the training condition were instructed that they would hear either a word or nonword, and their task was to decide which they heard for each trial. Once the participant had read the task instructions, they clicked a button on their web browser to begin the task. Listeners were asked to complete the task in one sitting and to make their response as quickly and accurately as possible. For each trial, listeners heard a token and were asked to indicate whether they heard a real-word or a non-word. Responses were made by pressing one of two keyboard keys ('z' = non-word, 'm' = real word). Once the listener submitted their response, playback of the next item began automatically (ISI = 1000 ms). Upon task completion, participants were taken to a separate page with instructions for the test phase of the experiment.

Test

All participants (both control and training groups) were randomly assigned to one of two test conditions: one group heard tokens from 7-step continua involving /θ/-/s/ minimal pairs

(e.g., *think - sink*), while the other group heard tokens from /θ/-f/ minimal-pair continua (e.g., *thought - fought*).

All participants were given instructions explaining the task and informed on the number of trials. They were asked to complete the task in one sitting, and to respond as quickly and accurately as possible. For each trial, participants heard a token and were asked to decide which of two words (presented textually on the screen) they had heard. Responses were again made via keyboard press ('z' = /θ/-word, 'm' = /f/- or /s/-word, depending on group). Textual response options were presented 500 ms prior to auditory stimulus onset (ISI = 1000 ms), following previous studies (Reinisch & Holt, 2014; Reinisch et al., 2013). In order to reduce response variability, six separate lists were created for each group, with each list blocked by continuum and order of continua and step number randomized. Participants were randomly assigned to one of these lists. Each step within a continuum was presented 4 times, making for a total of 112 tokens per list (4 continua x 7 steps x 4 repetitions).

Questionnaire

After completing the categorization task, participants were asked to complete a questionnaire with basic demographic information (age, race, gender, languages spoken, audio speaker type used to listen to stimuli, and experience with foreign-accented speech (from 1—7)).

Data preparation

Prior to statistical analysis, data were first processed by removing subjects who failed to meet experimental criteria. Of the 189 subjects initially recruited, 34 were removed based on performance in the training task: 33 subjects for failure to reach a 70% accuracy threshold on the exposure task, following Kraljic and Samuel (2006), 1 subject for categorizing over 50% of ambiguous /θ/ items as non-words (Norris et al., 2003). An additional 4 subjects were removed for failing to complete the test (categorization) task. Several subjects (N = 5) who had only a small number of missing trials (<5) due to a technical error were retained. An additional 10 subjects removed for failure to complete the questionnaire, and 2 subjects for indicating they were not native speakers of English. Subjects who had over 20% of trials with excessively quick responses (RT < 200 ms) were also removed (N=2) (following Reinisch and Holt, 2014, but using a less stringent exclusion criterion to minimize data loss). Finally, a further 17 subjects were discarded due to unwillingness or inability to reliably perceive the difference between continuum endpoints in the test task. Similar to Zheng and Samuel, 2020, we calculated a difference score by taking the per-subject proportion of [θ] responses for step 1 of each continuum (unambiguous [θ]), and then subtracting the proportion of [θ] responses at step 7 (unambiguous [s] or [f], depending on test condition). Based on visual inspection of the data, we set the difference score exclusion threshold at 20%. This is less stringent than the threshold used by Zheng and Samuel (2020), but reflects the greater perceptual similarity of the [θ] and [f] endpoints used in the current study, and was used primarily to exclude participants who were not performing the task in good faith. A total

of 69 subjects were thus excluded, and data from the remaining 120 subjects was further processed to remove trials with very fast (<200 ms) or very slow RTs (>2500 ms), following Reinisch and Holt (2014). This resulted in exclusion of 6.28% of trials in the lexical decision task and 5.27% of trials in the categorization task.

Analysis

Data were analyzed via mixed-effects logistic regression modeling using the lme4 package (Bates et al., 2014) in R (R Core Team, 2022). Maximal random effect structure was used for each model where this did not result in convergence issues, with random slopes fitted for all within-item predictors (Barr, 2013). The inclusion of fixed effects and interactions between them was decided via a step-wise selection process using a likelihood ratio test – terms that did not significantly improve model fit were removed.

Categorical variables were dummy-coded and included condition (training vs. control) and target word position (word-initial or word-final). Continuum step (1-7) was also included as a centered numeric variable (such that the reference level corresponded to the middle of the continuum). Random effects included by-listener intercepts and slopes for condition and for continuum step. Issues with model convergence were addressed by first simplifying random effect structure and then by simplifying interaction terms.

Results - Experiment 1

Lexical decision

Participants showed a high accuracy rate for all stimuli types, including the ambiguous /θ/ items types (see Table 2.1 for a summary of RTs and proportion items correct by stimulus type). RTs were somewhat higher for the non-word fillers compared to the other stimulus types, but there appeared to be no difference between filler words and critical /θ/ items, suggesting that the ambiguous pronunciation sounded relatively natural.

Table 2.1: Experiment 1 lexical decision task results: mean accuracy rates and response times (in milliseconds) for correct items, by word type.

	% Correct	RT (ms)
filler nonwords	94.2	1788
filler words	94.5	1646
critical /s/-words	97.4	1640
critical /θ/-words	93.2	1685

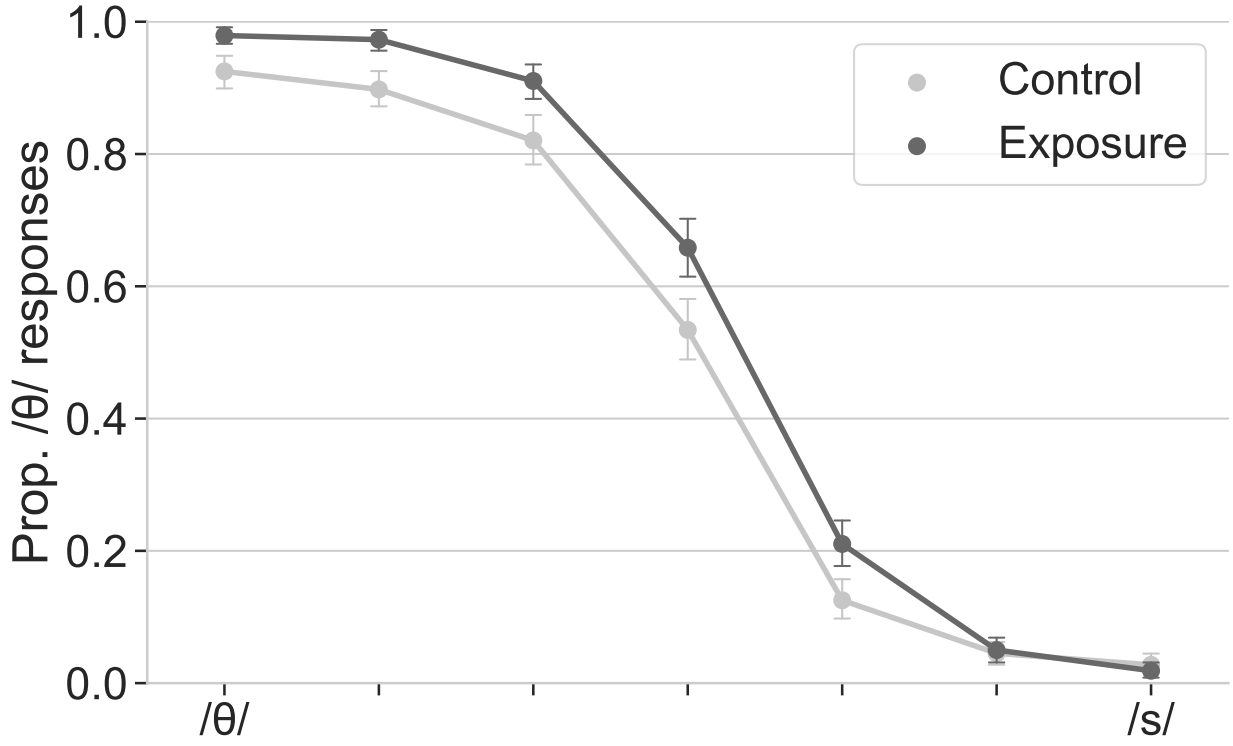


Figure 2.3: Proportion /θ/ responses for groups tested on categorizing 7-step [θ]-[s] phonetic continua, by exposure condition.

Categorization

Results of the categorization task for the /θ/-/s/ test group showed a significant effect of condition ($\chi^2(1) = 4.14$, $p < 0.05$), with participants exposed to ambiguous /θ/ = [θ/s] categorizing an average of 6.11% more tokens as /θ/ compared to controls (see Figure 2.3). There was also a significant effect of word position ($\chi^2(1) = 21.59$, $p < 0.001$), with word-initial [θ/s] targets showing a higher proportion of /θ/ responses overall (54.18%) compared to word-final tokens (48.34%). However, although the size of the training effect was larger with word-final continua (8.43% more /θ/ responses) than with word-initial ones (3.82%), there was no significant interaction of group and word position ($\chi^2(1) = 1.35$, $p = 0.24$), so it is unclear if word position affected the degree of perceptual learning (see Appendix A for a visualization of the recalibration effect by word position). Continuum step was also significant, with likelihood of a /θ/ response decreasing with a step number ($\chi^2(1) = 278.49$, $p < 0.001$), and there was a significant interaction of step and word position ($\chi^2(1) = 12.84$, $p < 0.001$), indicating that the effect of step was weaker with word-initial /θ/ targets.

Results for the [θ]-[f] group showed a small difference (1.33%) in the mean proportion of /θ/ responses between exposure and control groups (see Figure 2.4), but this was not

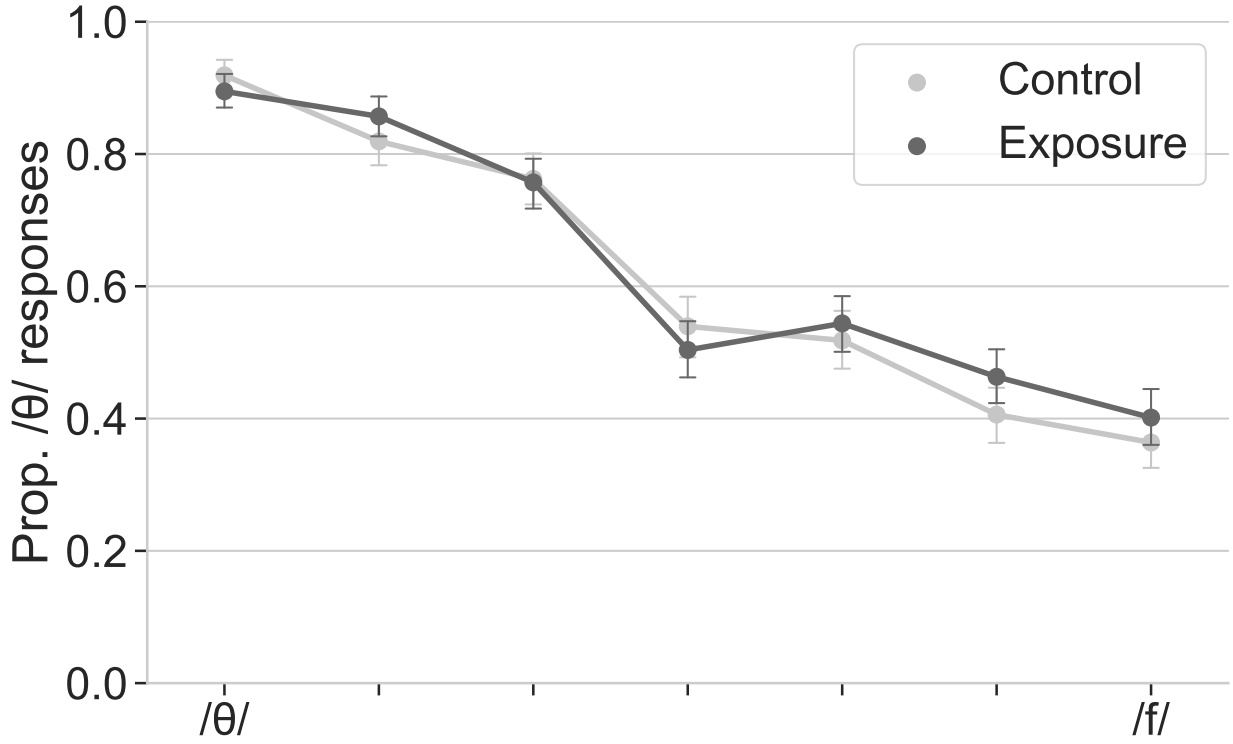


Figure 2.4: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [f] phonetic continua, by exposure condition.

significant ($\chi^2(1) = 0.01$, $p = 0.93$), suggesting that phonetic recalibration did not occur in this test condition. As in the [θ]-[s] condition, model results showed a significant effect of step ($\chi^2(1) = 430.11$, $p < 0.001$). There was also a significant effect of word position ($\chi^2(1) = 6.07$, $p < 0.05$), with a lower likelihood of /θ/ responses predicted for word-initial words – the reverse of the pattern shown for the [θ] – [s] test group. Results showed that listeners categorized 64.92% of tokens in the word-final continua as /θ/, but only 60.11% for word-initial continua.

Again, there was a small difference in proportion of /θ/ responses by word position across conditions: trained listeners showed 2.34% more /θ/ responses for word-initial tokens versus controls, and 0.39% more for word-final ones. There was no significant interaction of word position and exposure condition, although word-initial continua showed a larger difference between trained listeners and controls (see Appendix A.).

Exp.1 Discussion

Results from experiment 1 are consistent with a non-uniform category expansion mechanism of phonetic recalibration. Participants trained on an ambiguous [θ/s] realization of the

phoneme /θ/ showed a significant increase in the proportion of /θ/ responses compared to controls. However, participants trained on [θ/s] and tested on a distinct continuum [θ]-[f] did not show such a shift, which suggests that the learning strategy is contrast-specific and does not result in perceptual changes in other, neighboring parts of phonetic space – there was no evidence for a general expansion of phoneme categories.

Many previous tests of generalization in phonetic recalibration have focused on whether listeners generalize learning to a novel speaker with the same pronunciation. In this case, by contrast, we are essentially testing for generalization of a novel *pronunciation* within the same *speaker*. We do this by looking at the possibility of generalization of learning to a novel phonetic contrast involving the trained category. Results suggest that learning is contrast-specific, with no transfer of learning for the distinct fricative contrast [θ]-[f].

Given evidence from experiment 1 for a non-uniform category expansion, and prior evidence (White & Aslin, 2011) that recalibration may generalize to phonetically similar segments, experiment 2 tests for an effect of exposure to ambiguous [θ/s] on listeners' categorization of a [θ]-[ʃ] continuum. The question addressed by experiment 2 is whether the effect seen in experiment 1 is contrast-specific or represents a category expansion that could have an impact on the perception of a sound [ʃ] that is acoustically similar to [s].

2.3 Experiment 2

Previous research has shown high acoustic similarity of [θ] and [f]: both are low-amplitude and characterized by relatively flat acoustic power spectra Jongman et al., 2000; Stevens, 1998. They have also been shown to be perceptually confusable (Cutler et al., 2004; G. A. Miller & Nicely, 1955),as illustrated in Figure 2.2 (left panel).Categorization results for [θ]-[f] in the present study support this, with the flat, linear categorization functions (Figure 2.4) indicating that listeners found these stimuli highly confusable — even at the [f] endpoint of the continuum, listener responses are near-chance. This contrasts sharply with results for the [θ]-[s] test groups, which display an unambiguous s-curve typical of categorical perception (Figure 2.3). It is thus plausible that generalization failed due to low acoustic similarity between [s] and [f], but that there could still be a change in categorization of other fricatives that were more acoustically similar to [s]. For instance, the sibilant [ʃ] shares with [s] a relatively high amplitude and a prominent spectral peak, Jongman et al., 2000; Stevens, 1998, and it is perceptually similar (see Figure 2.2, left panel). While the results from experiment 1 are consistent with the non-uniform boundary expansion hypothesis proposed in the current study, they do not conclusively answer the question of whether learning is constrast-specific or not. Testing another portion of phonetic space proximal to [θ] but more acoustically similar to [s] could thus shed additional light on the nature of the recalibration mechanism.

Participants

An additional 112 participants were initially recruited via Mturk following the same inclusion criteria as in experiment 1. None had participated in experiment 1. Exclusion of participants who failed to meet experimental criteria left a total of 68 participants whose data were retained for data analysis, see section 2.3 for an explication of exclusion criteria). Participants had a similar demographic profile to those in experiment 1. They were primarily male (M = 42, F = 25), with 1 subject declining to report and no subjects selecting the 'other' option. Most participants self-reported as white (N=51), with 3 self-reporting as Asian, 6 as black, 3 as Hispanic, and 5 declining to answer. Participants spanned a wide range of age groups (1 = 18-20, 6 = 21-25, 11 = 26-30, 21 = 31-35, 13 = 36-40, 6 = 41 – 45, 4 = 46 – 50, 4 = 51-55, 2 = 61+). Participants self-reported moderate experience with accented English on a 7-point scale (M = 4.97, SD = 1.73, range = 1-7).

Stimuli

The same training materials used in the lexical decision task in experiment 1 were also used here. Each subject group was then tested on 4 separate minimal pair continua involving the /θ/-/ʃ/ contrast, e.g., *math* - *mash* (see Appendix A for a full list of test materials). The same procedure was used to morph naturally-produced recordings in 7-step intervals.

Procedure

The same training/test procedure used in experiment 1 was used, with the presence of a perceptual learning effect in the categorization phase tested by comparing an exposure group to a corresponding control with no training. Training items were identical to those used in experiment 1, and the number of tokens in the test phase also matched that of experiment 1.

Data preparation

Data from the 112 participants initially recruited was screened following the same criteria used in experiment 1. First, 33 subjects were removed for failure to meet the 70% accuracy criterion or to classify over 50% of the /θ/ target words as real words in the training task. An additional 2 participants were removed because they had already participated in the previous experiment. 1 subject was excluded for failure to complete the test task, 3 were excluded for failure to complete the questionnaire, and 1 for indicating that they were not a native speaker of English. An additional 2 subjects were excluded because over 20% of their trials had excessively fast reaction times (<200 ms). Finally, 2 subjects were excluded for failure to meet the difference score threshold, indicating that they could not reliably perceive the continuum endpoints. A total of 44 subjects were thus excluded from analysis. As in experiment 1, data from the remaining 68 subjects was filtered to exclude trials with

Table 2.2: Experiment 2 lexical decision task results: mean accuracy rates and response times (in milliseconds) for correct items, by item type.

	% Correct	RT (ms)
filler nonwords	90.0	1780
filler words	91.6	1640
critical /s/-words	92.4	1658
critical /θ/ words	90.7	1798

excessively fast (<200 ms) or slow (>2500 ms) reaction times. This resulted in the removal of 10.9% of trials for the categorization task.

Analysis

Model selection procedures, coding scheme for predictor variables, and random effect structue were all identical to experiment 1.

Exp.2 Results

Lexical Decision

As in experiment 1, participants showed a high accuracy rate for all stimuli types, including the ambiguous /θ/ items types (see Table 2.2). Unlike in experiment 1, however, critical /θ/ items appeared to pattern with filler nonwords, showing higher RTs than for filler words /s/-words. Accuracy rates for critical /θ/ words were nonetheless comparable to other items, indicating that participants accepted these items as sufficiently natural.

Categorization

Analysis of the categorization data showed a significant effect of condition ($\chi^2(1) = 5.33$, $p < 0.05$), with participants in the exposure group classifying an average of 7.08% more tokens as /θ/ compared to controls (see Figure 2.5). Participants showed almost the same size training effect for word-initial continua (7.37% more /θ/ responses vs. controls) as word-final ones (6.79% more /θ/ responses vs. controls), with no significant interaction between word position and exposure condition ($\chi^2(1) = 0.31$, $p = 0.58$). Again, a visualization of the training effect by word position can be found in Appendix A. Results also showed a significant effect of continuum step ($\chi^2(1) = 231.68$, $p < 0.001$), with a significant interaction of step and target word position ($\chi^2(1) = 23.51$, $p < 0.001$) as well as step and condition ($\chi^2(1) = 7.38$, $p < 0.01$).

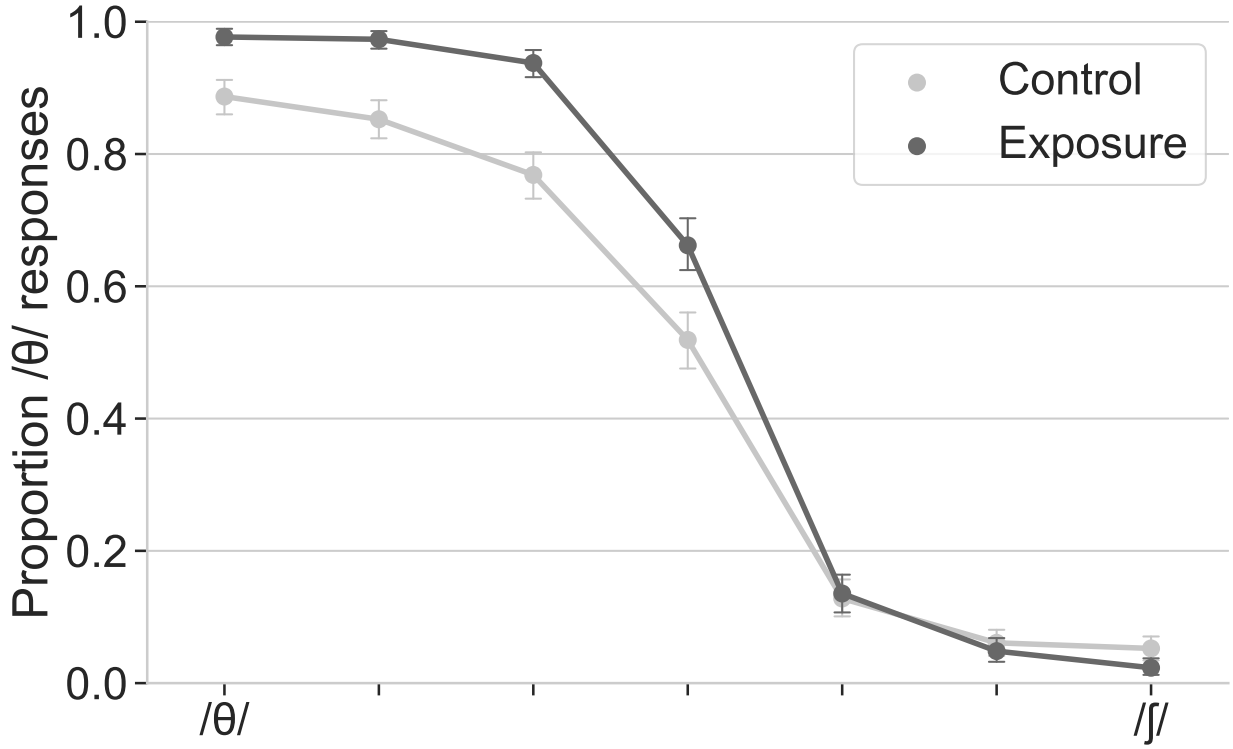


Figure 2.5: Proportion /θ/ responses for groups tested on categorizing a 7-step [θ] – [ʃ][phonetic continuum, by exposure condition.

Exp.2 Discussion

Results of experiment 2 showed that listeners exposed to an accent involving /θ/ = [θ/s] generalized learning to neighboring phonetic sounds, as shown by the clear categorization shift observed with [θ]-[ʃ] minimal pairs. These results build upon what was found in experiment 1, where learning appeared to be specific to the trained accent, with a categorization shift observed with [θ]-[s] minimal-pair continua, but not with [θ]-[f]. They illustrate that generalization of learning beyond that specific accent presented in exposure is possible, provided that the test sounds are phonetically similar enough. Thus, recalibration can result in changes to perception of neighboring phonetic categories when those are perceived to be sufficiently similar to sounds used in the exposure accent.These results thus add to existing recalibration literature which has found generalization of learning to novel phonetic contrasts Kraljic and Samuel, 2006; Schuhmann, 2014.

2.4 General Discussion

The goal of the current study was to investigate the perceptual mechanism underlying phonetic recalibration. In particular, we wanted to test whether the change in categorization behavior following exposure to an atypical pronunciation of a particular phoneme was due to a targeted shift in the phonetic category, or whether it was due to a more general mechanism of relaxing phonetic categorization criteria or expanding the category into neighboring perceptual space. Results of the current study provide support for a category expansion strategy, albeit one that is sensitive to the phonetic similarity between the exposure accent and neighboring sounds consistent with the non-uniform category expansion proposed here (schematized in panel C of Figure 2.1).

The fact that perceptual learning in the present study generalized to a neighboring phonetic contrast involving the same target category /θ/ used in the exposure phase suggests that a simple boundary shift is not the underlying mechanism at work in recalibration—there must be some degree of category expansion, affecting subsequent perception of neighboring sounds. However, listeners cannot be doing this willy-nilly. That is, the strategy cannot simply be "this speaker has an unusual /θ/ pronunciation". If this were the case, we would expect any atypical realization of /θ/ to be acceptable as an instance of this phoneme, but instead we see that the changes in perception following exposure to the accent appear to be constrained by (perceived) phonetic distance from the exposure accent. Listeners seem to be relying on some dimension of phonetic similarity to limit the degree of category expansion. In this case, it is possible that the spectral characteristics of [ʃ] and [s] are close enough for listeners, with the result that both categories show more overlap with neighboring /θ/. Meanwhile, because [f] is acoustically (Jongman et al., 2000; Stevens, 1998) and perceptually (Cutler et al., 2004; G. A. Miller & Nicely, 1955) distinct from [s], it appears to be unaffected by recalibration.

These results align with a large body of work on phonetic recalibration and perceptual learning more generally which suggests that phonetic similarity between exposure and test sounds plays a critical role in generalization of learning. For instance, studies by Kraljic and Samuel (2005) and Eisner and McQueen (2005) have shown that generalization of ambiguous fricative pronounciations [s/f] or [s/ʃ] to new speakers is highly constrained. A possible explanation for this (Kraljic & Samuel, 2007) is that fricatives encode speaker-specific spectral information, and listeners are sensitive to acoustic differences in the way that these sounds are realized. More recent work bolsters this hypothesis, showing that when such differences between exposure and test speakers are minimized, generalization is possible. Reinisch and Holt (2014), for instance, found that listeners exposed to an ambiguous [f/s] accent in a Dutch-accented English speaker generalized to a novel Dutch-accented speaker when exposure and test speakers' fricatives were sampled from a similar perceptual space. Finally, an accent accommodation task by Xie and Myers (2017) found that successful transfer of learning for accented speech depended on the acoustic similarity between exposure and generalization speakers. Listeners did not appear to be learning an accent-wide schema that could be applied to new speakers, but rather more specific characteristics of the target sounds

they were tested on.

The finding of generalization in the current study from a /θ/ = [θ/s] pronunciation to the /θ/-/ʃ/ contrast is particularly interesting as most previous work suggested that generalization of perceptual learning for fricatives is limited (Eisner & McQueen, 2005; Kraljic & Samuel, 2005) and has been found only when perceptual similarity between exposure and test speakers is controlled for (Reinisch & Holt, 2014). If, as earlier work has suggested, the specificity of learning in these cases has to do with listener sensitivity to the different acoustic realization of fricatives by male and female speakers, it is surprising to see generalization in the current study, where the differences between training and test contexts may be of similar magnitude. In the current study, the transfer of learning from an /s/-like realization of /θ/ to an /ʃ/-like one suggests that perceptual learning, at least for some fricatives, may be coarser-grained. Importantly, however, within-speaker transfer of learning across categories could be a fundamentally different process than the cross-speaker generalization tested in earlier work, so comparisons between the two should be made cautiously.

As we have already noted, the finding of generalization of perceptual learning found in the present is not novel on its own. The key finding in the current study is that learning can generalize to a novel phonetic contrast involving the same underlying category manipulated in training, a scenario that has received little attention in prior work. Generalization observed in previous recalibration studies typically involves sound contrasts where both categories are distinct from the trained phoneme, e.g., where listeners are trained on /d/ = [d/t] and generalize to the [b]-[p] contrast (Kraljic & Samuel, 2006, 2007; Mitterer et al., 2016; Mitterer & Reinisch, 2017; Schuhmann, 2014). Closer to the current study, generalization has also been found across languages for bilingual speakers, where learning-induced categorization shifts transfer from L1 to L2 or vice versa (Reinisch & Holt, 2014; Schuhmann, 2016). However, as the authors note, differences between each of the trained sounds /s/ and /f/ across these languages (Dutch, German, and English) are minimal (especially if speakers produce the L2 versions with an L1 accent). More relevant would be an example of transfer across languages with more distinct realizations of a given category (e.g., Spanish dental /d/ vs. English alveolar /d/), but we are not aware of work that has tested this.

As we have seen, existing evidence for a category expansion account of perceptual learning across different tasks is mixed. In some cases, what we see looks more like an indiscriminate relaxation of categorization criteria that may extend beyond the trained category or phonetic pattern. In the face of uncertainty about the input or extensive variation, the perceptual system might adopt an "anything goes" adaptive strategy. However, the current study does not support such a view. While results clearly show that changes in categorization criteria can extend beyond the exact phonetic pattern in exposure stimuli, they also show that these changes are constrained by similarity to the trained accent. Results of the current study thus provide support for (a version of) the category expansion hypothesis discussed in earlier work (Kleinschmidt & Jaeger, 2015; Schmale et al., 2012).

However, additional research is necessary to show that this strategy is not specific to the particular sound categories and contrasts used in these experiments. Earlier work suggests that recalibration strategies may differ depending on the sounds involved (Kraljic & Samuel,

2006, 2007). Crucially, we have assumed in the current study that learning occurs at the level of the trained category (in this case, /θ/). However, other studies indicate that learning may occur at a more abstract (e.g., feature) level. For instance, Kraljic and Samuel (2006) have found that listeners can generalize a stop voicing contrast across place of articulation, while Schuhmann (2014) found generalization of a fricative pronunciation involving [f/s] to their voiced counterparts [v]-[z]. So, it is plausible that listeners in the current study are abstracting over acoustic differences between [s] and [ʃ], and instead focusing on their shared properties as sibilants — relatively loud fricatives with spectral similarities.

Such results have implications for the learning of natural accents. Recent work on phonetic recalibration has investigated whether this mechanism is responsible for perceptual adaptation to non-native accents (Reinisch & Holt, 2014; Zheng & Samuel, 2020), and this remains an open question in the literature. While the present study does not directly address this question, results suggest that the perceptual mechanism underlying the recalibration effect is flexible enough to accommodate differences between exposure and generalization contexts. This is important, because to adapt to a natural accent, listeners must be able to handle the substantial variation that exists within speakers of a given accent group, where the same target L2 category may be realized in multiple distinct ways (Seibert, 2011). Thus, it can behoove listeners to generalize to some extent or to maintain a scope of learning that is relatively coarse. In other words, it is plausible that recalibration is not based on the precise acoustic properties of heard exemplars, but rather on rougher measures of phonetic similarity. Taken together, these results support a view of the perceptual system as flexible and able to adjust to variation in the environment, but simultaneously constrained to not overgeneralize learning.

Chapter 3

Effects of phonetic learning on online lexical processing

3.1 Introduction

Previous research on perceptual learning for speech has employed a wide variety of measures to assess the effects of perceptual learning on subsequent language processing. Such measures range from comprehension-focused tasks such as word/sentence transcription in noise (Baese-Berk et al., 2013; Bradlow & Bent, 2008; Melguy & Johnson, 2021; Vaughn, 2019) to tasks measuring lexical processing fluency (Clarke & Garrett, 2004; Xie, Weatherholtz, et al., 2018) to tasks focused on measuring changes in phonetic category boundaries (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Norris et al., 2003).

Crucially, it has often been assumed in the literature that accent accommodation is driven by lexically-guided changes in phonetic category structure (e.g., (Bradlow & Bent, 2008)). This explanation is intuitively appealing: if listeners "re-tune" their phonetic category boundaries following exposure to a speaker with an atypical pronunciation of a given sound, then future encounters with this pronunciation should be less surprising for the listener and incur reduced processing costs. Since it has been shown that even minor (subphonemic) changes to the way a given sound is realized can incur processing costs (Whalen, 1984), it is plausible that making the appropriate adjustements could reduce these costs by eliminating mismatches between listener expectations and a given speaker's pronunciation. Since accented speech typically differs from natively-accented speech in systematic ways due to the influence of the L2 learner's L1 phonological system, there are regularities within and across L2 speakers of a given L1 background which listeners can exploit. For instance, Japanese speakers typically struggle with the English /r/-/l/ contrast, substituting their Japanese /r/ which is phonetically intermediate to the two English sounds. If listeners can adapt to this regularity by adjusting their expectations about what types of pronunciations are likely L2 realizations of a given native phoneme or allophone, then this could faciliate perception of any speaker who shares this accent. However, as has been pointed out in recent

literature (e.g., Zheng and Samuel, 2020) there is limited evidence for a direct relationship between recalibration and accent accommodation, and existing studies that have sought to address this question explicitly yield mixed results (Charoy, 2021; Reinisch & Holt, 2014; Xie et al., 2017; Zheng & Samuel, 2020).

The primary goal of this chapter is to investigate whether changes in category structure indeed result in improvements to processing of accented speech. If this relationship exists, it would suggest that phonetic category recalibration may indeed be a driving mechanism for perceptual accommodation to an unfamiliar accent. The set of experiments in Chapter 2 provided evidence for a limited form of the category expansion hypothesis as a mechanism for phonetic recalibration following exposure to an atypical pronunciation. Results showed that exposure to an ambiguous pronunciation /θ/ = [θ/s] affected subsequent perception of both /θ/-/s/ and /θ/-/ʃ/ minimal-pair phonetic continua. If the same mechanism that drives recalibration of phonetic category boundaries also affects online lexical processing of that accent, then we should observe a relationship between category shifts and improved processing for accented words. Specifically, we expect listeners to show improved processing not only for novel words produced with the exposure accent, but also for words produced with a distinct but perceptually similar accent, consistent with a category expansion mechanism.

3.2 Background

Effects of perceptual learning on word processing

A relatively large body of literature has utilized lexical processing as a measure of accent accommodation, often with a focus on segment-specific adaptation. If recalibration of category boundaries coincides with improvement in the accuracy and/or fluency of lexical processing, then it is plausible that there is a link between natural accent accommodation and category boundary recalibration. Previous research has found that accent exposure often does result in improved lexical processing. A seminal study by Clarke and Garrett (2004) showed that listeners can begin to see such benefits in under a minute of exposure. In their study, they exposed listeners to 3 blocks (each 12-18 sentences) of Spanish- or Mandarin Chinese-accented English where the final word in the sentence was unpredictable, and listeners' task was to indicate if a visual target on the screen matched the spoken word or not. Participants saw a steady decrease in response times, and by the fourth block these approached the baseline established for native (unaccented) speech. By contrast, a control group that performed the same task but heard unaccented sentences in the first three blocks saw a sudden increase in response times when tested on accented sentences in the fourth block. A recent study by Xie, Weatherholtz, et al. (2018) successfully replicated this basic effect with the same (Mandarin Chinese) accent, finding also that learning could transfer to a novel Mandarin-accented speaker provided there was sufficient acoustic similarity between speakers. This suggests that listeners learned an accent-wide schema that could generalize to speakers sharing the same pronunciation.

A number of studies utilizing artificial accents have used different measures to assess the effect of perceptual learning on lexical processing. For instance, Maye et al. (2008) exposed listeners to an artificial accent where all front vowels were lowered (e.g, *witch* → *wetch*), followed by a test phase where participants had to decide whether each word was a real word or non-word (lexical decision task). The authors found that listeners trained on the accent were more likely to accept novel words with lowered front vowels as real words, but not words with raised front vowels (a different pattern than they had been exposed to). A study by Weatherholtz (2015) assessed perceptual learning of a similar accent which mirrored an actual vowel-shifted pronunciation in American English dialects. Like Maye et al. (2008), this study used proportion of 'word' responses in an auditory lexical decision task to assess learning, as well as a naming task to assess listener recognition of accented words (accuracy and response latency). Overall results showed generalization of learning to a novel accent where back vowels were raised instead of lowered (contra Maye et al., 2008), as well as to a structurally parallel pattern where front vowels were lowered. However, listeners who were exposed to a different accent containing a pattern of back vowel raising showed no generalization to raised front vowels. Another study by Cooper and Bradlow (2018) exposed listeners to an artifical accent where target vowels and consonants were substituted with other sounds to create a non-standard English accent (e.g., *cream* → *crim*, *throw* → *trow*). To evaluate perceptual learning, they used both an auditory lexical decision task and a word identification task where participants were asked to transcribe the word (denoting non-words with an 'X'). Results from both tasks showed that listeners previously exposed to the accent were more likely to accept novel words containing the same accent as real words. There was also an overall increase in trained listeners' willingness to accept any atypical tokens as real words, even when the pronunciation did not match the exposure accent. Together, these studies suggest that listeners may be using a form of category expansion to adapt to an artificial accent.

Other studies have used a cross-modal priming paradigm to evaluate the effects of perceptual learning on lexical processing (Charoy, 2021; Eisner et al., 2013; McQueen et al., 2006; Witteman et al., 2013; Xie et al., 2017). This is a visual lexical decision task where participants hear an auditory prime and must decide whether a written word on the screen is a real word or not. Critical trials exploit the fact that the strongest priming effects in such a task are observed when the auditory prime matches the written target (identity priming). Thus, the size of the priming effect can be interpreted as a measure of the degree of lexical activation: if there is a training effect, we expect stronger lexical activation for trials with accented 'identity' primes, suggesting improved processing of the accent. This kind of methodology has been used to evaluate perceptual learning for both naturally-accented speech and for experimentally-created artificial accents. For instance, Eisner et al. (2013) tested British English listeners on a word-final stop devoicing pattern found in Dutch-accented English. They found that listeners were able to adapt to the pronunciation, whether it was produced in a natural accent by an L2 English (L1 Dutch) speaker, or as an artificial accent where a British English speaker imitated the accent (e.g., where *overload* → *overloat*). Trained listeners showed a stronger priming effect in trials containing devoiced primes (e.g., *seat*

primed both written <seed> and <seat>). Witteman et al. (2013) investigated the effect of familiarity with German-accented Dutch on perceptual learning of this accent. They used a cross-modal priming task to measure fluency of word recognition for words produced with a weak, medium, or strong German accent (strongly-accented primes contained relatively salient substitutions of Dutch vowels, medium-accented primes vowels contained less salient vowel substitutions, and weakly-accented primes contained no substitutions). Initial results showed that participants with limited prior accent experience adapted to weakly- and medium-accented primes, but not strongly-accented ones, whereas participants with extensive prior accent experience could adapt to all 3 types. A followup experiment exposed listeners with limited previous accent experience to the same speaker as in the test phase. They found that this 4-minute exposure phase was sufficient to produce reliable priming effects for strongly accented items that were on par with those seen with weakly- and medium-accented primes. Listeners exposed to the same accent but without the strongly-accented items also saw a benefit — they were able to adapt to the strongly-accented items by the second half of the experiment. Such studies indicate that accent accommodation results in improvements to lexical processing when accented words are sufficiently similar to the training accent.

Phonetic category recalibration and accent accommodation

Several recent studies have explicitly set out to answer the question of whether accent adaptation (measured as changes in lexical processing) is indeed driven by changes in phonetic category structure.

For instance, Reinisch and Holt (2014) addressed this question by testing whether phonetic recalibration could be observed in the context of speech that was already (globally) accented. Previous literature had only tested recalibration in the context of an otherwise natively-accented speaker with just a single target sound manipulated to be ambiguous. The authors reasoned that if recalibration were a plausible mechanism for listener adaptation to accented speech, it should occur even when a speaker already deviates from native-speaker standards on multiple phonetic dimensions. They exposed American English listeners to Dutch-accented English where a single sound (/s/ or /f/) had been manipulated to be phonetically ambiguous ([s/f]). They then tested listeners on categorization of /s/ - /f/ minimal-pair continua (e.g., *nice* - *knife*). They found that listeners shifted their category boundary for both the exposure speaker and for novel speakers with the same Dutch accent, provided those novel speakers' stimuli were acoustically similar to that of the training speaker.

Xie et al. (2017) provided a more explicit test of this question by evaluating the effect of accent exposure on multiple measures of phonetic category structure. They exposed American English listeners to Mandarin-accented English where word-final /d/ was produced in a way that was perceptually ambiguous between /d/ and /t/ for listeners not familiar with the accent (the speaker's production of this sound spanned the range from unambiguous /d/ to unambiguous /t/). Following accent exposure, they tested listeners on (1) a cross-modal priming task involving critical word-final /d/ primes, (2) a two-alternative forced choice

categorization task with word-final /d/ or /t/ minimal pairs, and (3) a rating task where listeners rated the goodness of /d/- and /t/-final words as exemplars of each respective category. Overall results showed that compared to untrained controls, trained listeners adjusted their /d/-/t/ category boundary (accepted more tokens as /d/), were more likely to process /d/-final words as containing /d/ (i.e., showed stronger identity priming with /d/-final primes), and showed higher goodness ratings for both /d/- and /t/-final words. This suggests that accent exposure results in multiple changes to internal category structure — not just shifting boundaries, as had been observed in previous work.

Zheng and Samuel (2020) tested a similar question using a different methodology. They investigated whether changes in category boundaries and accent accommodation are driven by the same underlying mechanisms. To do this they had American English listeners complete a standard recalibration task involving an ambiguous /θ/= [θ/s] pronunciation in an otherwise natively-accented American English speaker. The same group of listeners also completed an accent accommodation task. This involved a brief period of exposure to Mandarin-accented English, followed by a task where listeners judged Mandarin-accented words (e.g., *think* produced as *sink*) as real words or non-words. Results showed that listeners showed both a reliable shift in their [θ]-[s] category boundary as well as an increase in proportion of accented critical words that were accepted as real words. Listeners also showed a general increase in the proportion of words accepted as real words, regardless of whether they matched the exposure accent. However, a correlational test found no relation between boundary shifts and accent accommodation, leading the authors to conclude that the two were dissociated and that accent accommodation was likely driven by a general relaxation of categorization criteria.

Charoy (2021) also investigated whether recalibration underlies accent accommodation. In contrast with the typical recalibration literature, her study looked at both "bad maps" (substitutions of a native phoneme with another, cf. Sumner, 2011) as well as ambiguous sounds typical of recalibration studies. Her results showed that recalibration of category boundaries can be observed for both ambiguous and remapped phonemes. However, subsequent experiments that focused on lexical processing as a measure of accent adaptation (cross-modal priming) found little evidence that these changes improved lexical processing for either type of accent. For instance, they found that "bad map" productions of /θ/ (as [s]) yielded no priming for /θ/ words for either trained or control listeners. Moreover, ambiguous /θ/ primes ([θ/s]) primed both /s/ and /θ/ words, with no difference found between listeners with prior exposure to the accent and controls. The sole effect of training was seen with a bad map production of /d/ as [t], where devoiced *fond* primed written <fond> for trained listeners but not controls, replicating the basic finding of Xie et al. (2017). However, listeners in this latter set of experiments showed no evidence for recalibration of category boundaries, contrary to the first experiment. The author concluded that a form of criteria relaxation was the more likely perceptual accommodation strategy for accent processing.

This set of studies yields inconclusive results on the relationship between category boundary shifts and lexical processing measures for accented speech. For instance, although Reinisch and Holt (2014) show that recalibration is possible for ambiguous target phonemes

in the context of an already accented speaker, this does not directly demonstrate a connection between accent accommodation and recalibration, but simply that recalibration is not blocked in this context (as previously noted by Zheng and Samuel, 2020). Zheng and Samuel (2020) notably fail to find a correlation between recalibration and accent accommodation. However, their study measured recalibration for an ambiguous /θ/ = [θ/s] accent in an otherwise American English accent, whereas the accent accommodation task utilized Mandarin-accented English. The use of different speakers across tasks makes it difficult to assess whether recalibration drives accent accommodation. Xie et al. (2017) provide the strongest evidence that recalibration yields improvements in processing, finding that exposure to a Mandarin-accented speaker with ambiguous word-final /d/ productions lead to shifts in /d/-/t/ category boundaries, improved category goodness ratings for exemplars of that speaker's /d/ and /t/, and improved lexical processing in a cross-modal priming task involving the critical phonemes. However, the methods used in this study deviate in several ways from the typical recalibation study. Most importantly, the authors used Mandarin-accented tokens of word-final /d/ which, as they note, spanned a phonetic range between unambigous /d/ and unambiguous /t/. Moreover, Charoy (2021) failed to replicate the basic findings of Xie et al (2017). Evidence for phoneme-specific adaptation in the cross-modal priming experiments in her study was very weak, with no effect of training found for most of the sound contrasts she tested. Moreover, the fact that no recalibration was observed in her final set of experiments makes it difficult to make any conclusions about the connection between boundary shifts and lexical processing. Finally, all of the studies reviewed here only tested a single accent — stimuli from the test phase closely matched those in the exposure phase. This contrasts with the approach of the experiments in the current study, which purposefully test listeners on both the same accent as in training, and on a similar but distinct accent involving the same critical phoneme. Providing both types of test may yield more insight on the nature of the mechanism that drives changes in lexical processing.

The current study

The experiments in this chapter explicitly test the question of whether changes in phonetic category representation facilitate subsequent lexical processing. If targeted adjustment of specific sound categories following accent exposure is a mechanism for accent adaptation, then we ought to see a connection between changes in phonetic category representation, and corresponding changes in online lexical processing. Specifically, we expect that phonetic recalibration will result in reduced processing difficulty for atypical productions of the accented target sound, leading to more rapid categorization of that sound as the intended category. Previous literature that has examined the relationship between category changes and lexical processing has yielded mixed results (Charoy, 2021; Xie et al., 2017), and the use of different methodologies across studies make it difficult to draw clear conclusions. Moreover, while there is limited support for a relationship between category changes and lexical processing, it is still unclear whether the same mechanisms are responsible for changes in both domains, because prior literature has not addressed this question directly. For instance, it is pos-

sible that listeners use a relatively specific mechanism (non-uniform/asymmetric category expansion) to recalibrate category boundaries, but that lexical processing is shaped by more general learning mechanisms. For instance, Cooper and Bradlow (2018) found that exposure to an artificial accent led to both pattern-specific changes in lexical processing, but also a broader tendency for listeners to accept non-words as real words, suggesting a uniform relaxation of categorization criteria. Given that Chapter 2 did not find evidence for such general category expansion, it is plausible that changes to internal category structure and lexical processing are (to some degree) independent.

Hypotheses

To test this question, the following experiments utilize a cross-modal priming task, following previous literature which has used such a task to evaluate the success of accent adaptation (e.g., Xie et al 2017, Charoy 2021). If changes in category structure affect subsequent lexical processing, listeners with prior exposure to the accented speaker should see stronger priming effects for critical words exemplifying the accent compared to controls. If the same mechanisms that constrain recalibration of category boundaries apply to online processing, then we should see such a facilitative effect of accent exposure for listeners tested on the same accent as in training (/θ/ = [θ/s], Experiment 3) as well as those tested on a similar, but distinct accent (/θ/ = [θ/ʃ], Experiment 4). If both of these findings align with the patterns of category recalibration seen in Chapter 2, this would provide evidence that changes to category structure result in improved lexical processing, and that the same mechanisms of phonetic learning apply across both domains. This result would support recalibration of category boundaries as a plausible mechanism for the more complex real-world task of adapting a non-natively accented speaker.

3.3 Experiment 3

Experimental platform

The experiments in this chapter were conducted using the same custom web-based program used previously in Experiments 1-2. However, in these and all subsequent experiments, data collection occurred through Prolific rather than Mturk. Prolific, like Mturk, is an online platform where workers are compensated for performing various tasks. However, recent studies suggest that Prolific yields higher-quality data (Peer et al., 2021), a finding corroborated in this study by the much lower participant exclusion rates compared to Experiments 1 and 2.

Participants

81 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 12 participants, leaving a total of 69

whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. An additional 19 subjects were recruited on Mturk under the same selection criteria to participate in a pretest, which was used to select critical /θ/ tokens to be used in the main experiment. Participants were primarily male (M = 38, F = 27), with a small number selecting the 'other' response option (N = 2). Participants self-reported the following ethnicities: white (N = 46); black (N = 7), hispanic (N = 6), asian (N = 5), and other (N = 3). They spanned the following age groups: 18-20 (N = 5), 21-25 (N = 16), 26-30 (N = 17), 31-35 (N = 5), 36-40 (N = 10), 41-45 (N = 4), 46-50 (N = 4), 51+ (N = 6). They reported moderate experience with accented English on a seven-point scale (M = 5.0, SD = 1.5, range = 2-7).

Stimuli

The same training materials used in Experiments 1-2 were used in this study. The same speaker also produced all test materials. Materials in the test phase consisted of 180 written targets (90 words + 90 nonwords), paired with 180 auditory primes (90 words + 90 nonwords). Auditory primes were constructed to result in three types of experimental conditions, evenly split across trials. In identity trials, the auditory prime matched the written target (e.g., spoken *banana* + written <banana>). In related trials, the prime was a minimal non-word phonologically related to the target, but mismatching in the first phoneme (e.g., *ganana* + <banana>). Finally, in unrelated trials, the auditory prime had no semantic or phonological relation to the target (e.g., *computer* + <banana>). Target words consisted of 60 critical words containing /θ/ and 120 filler words and nonwords (30 words and 90 nonwords). Like fillers, critical targets were evenly split to contain either identity, related, or unrelated primes. Identity primes in this case were words containing ambiguous /θ/ = [θ/s] (e.g., *?erapy* + <therapy>). Related primes always contained an unambiguous /s/ (e.g., *serapy* + <therapy>). The 60 written /θ/ target words were divided into 3 lists of equal length, which were equated in mean lexical frequency (9.4, 9.8, and 9.4 occurrences per million in the SUBTLEX corpus, respectively) and mean word length (each list averaged 2.2 syllables per per word, range 2-4). For each written target word, a corresponding identity, related, or unrelated prime was recorded, resulting in a total of 180 auditory primes for the critical trials. Primes were counterbalanced across the 3 /θ/ target word lists. This resulted in a total of 6 counterbalanced lists, each of which had the same 60 written /θ/ targets. As in Experiments 1-2, critical primes with ambiguous /θ/ = [θ/s] were created using the same morphing procedure that was used to create the ambiguous /θ/ words in the training task, with a separate norming task determining selection of the most ambiguous token. Finally, both test and exposure items were normalized in amplitude, resampled to 16 kHz, and converted to mp3 format to ensure between-browser compatibility during audio presentation.

Procedure

Pretest

To select the most ambiguous critical /θ/ items for the exposure task, the 60 /θ/-/s/ continua were subjected to a pretest by a separate listener group. For this purpose, 19 participants were initially recruited via MTurk. Data from 3 participants were removed due to inability or unwillingness to perceive the difference between items (<30% mean difference in responses between continuum endpoints). The procedure closely matched the norming task used in Experiment 1. For each 7-step continuum, the endpoints were left out, leaving the middle five steps (2-6). Stimulus presentation was blocked by continuum, and 7 lists were created, each with a different random order of continua and step numbers with a given continuum. Participants were randomly assigned to one of these lists. Each participant was thus asked to judge a total of 300 tokens (60 continua x 5 tokens each). Once the participant had read the task instructions, they clicked a button on their web browser to begin the task. Listeners were asked to complete the task in one sitting and to make their response as quickly and accurately as possible. For each trial, listeners heard a token and were asked to indicate whether they heard a real word (e.g., therapy) or a non-word where the /θ/ had been replaced with /s/ (e.g., *serapy*). Responses were made by pressing one of two keyboard keys ("z" for non-word, "m" for real word). The response labels were present on the screen for the duration of the whole trial. Once the listener submitted their response, playback of the next item began automatically [inter-stimulus interval (ISI) = 1000 ms]. The selected ambiguous step number for each continuum was based on the step closest to the category boundary between response options (i.e., the point at which 50% of participants indicated hearing the /θ/-word). In cases where the percentage of /θ/-word responses for this token fell below 50%, the previous step on the continuum was selected. This was done to compensate for the real-word bias in speech perception (Ganong, 1980) and to ensure that the target phoneme was phonetically ambiguous, following Reinisch et al. (2013). Results for the selected step closely matched those obtained in their study: the average step number of the ambiguous token was near the middle of the seven-step continuum (4.40), and the mean percentage of /θ/-word responses at this step was 70%.

Screening

Prior to proceeding to the experimental task, all participants completed a screening task designed to ensure that they were wearing headphones, identical to that used in Experiments 1-2. Questionnaire results again suggested that the task was highly effective, with the majority of participants self-reporting use of high-quality headphones (N = 49) or ear-buds (N = 18)

Lexical decision task (exposure)

Participants were randomly assigned to a training or control group, both of which completed a lexical decision task designed to familiarize listeners with the speaker's voice. Stimuli and procedure for the training group were identical to those used in Experiments 1-2. Participants in the control group completed the same task, except that critical words containing ambiguous /θ/ were replaced with filler words matched on mean syllable count and lexical frequency.

Cross-modal priming task (test)

Following the lexical decision task, all participants completed a cross-modal priming task. Participants were given instructions explaining the task and informed on the number of trials. They were asked to complete the task in one sitting and to respond as quickly and accurately as possible. For each trial, participants heard an auditory prime, followed by the presentation of a written target word on the screen, and were asked to indicate whether the written word was a word or non-word. Responses were again made via keyboard press ("z" = non-word, "m" = real word). For each trial, the written target appeared immediately upon prime offset (Charoy, 2021; Xie & Myers, 2017; Xie et al., 2017), with an inter-trial interval of 1000 ms. Listeners in both control and training groups were randomly assigned one of the 6 counterbalanced test lists described above, each containing a total of 180 trials.

Questionnaire

After completing the categorization task, participants were asked to complete the same questionnaire used in previous experiments, containing questions about audio hardware used and participant demographics [age, race, gender, languages spoken, type of speaker used to listen to stimuli, and experience with foreign-accented speech (from 1 to 7)].

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 5 participants were removed for failing to complete the questionnaire, 3 for self-reporting as non-native speakers of English, 1 for failing to complete the exposure (lexical decision) task, 3 for failing to meet the 70% accuracy threshold in the test (cross-modal priming) task, and 2 for responding too quickly (RT $<$ 200 ms) on over 20% of trials, following the criterion used in Experiments 1-2. An additional 3 participants were excluding for restarting the experiment midway through. Resulting data were subsequently cleaned to remove responses that were too fast (RT $<$200 ms) or too slow (RT $>$ 2500 ms), following the procedure in Exp. 1-2. This resulted in the removal of 184 trials (1.6%) of data. Finally, incorrect responses were removed prior to analysis (495 trials or 4.4% of data). This left a total of 10,840 observations (from 64 participants) that were retained for the analysis.

Analysis

Data were analyzed via mixed-effects linear regression modeling using the lme4 package (Bates et al., 2014) in R (R Core Team, 2022). Random intercepts were fitted for participant and target word, and by-participant random slopes were fitted for all predictors where this was justified by experimental design (Barr, 2013). Only critical trials were included in the analysis. Reaction times were filtered to remove incorrect responses prior to analysis, then converted to a log scale and normalized (converted to z-scores). Predictor variables were treatment-coded and included experimental group (training vs. control, ref = control), condition (identity, related, or unrelated prime, ref = unrelated). Given that previous cross-modal priming studies have observed perceptual learning over the course of the experiment, the analysis also included experiment half (block 1 vs. block 2, ref = block 1), following Witteman et al. (2013). Significance of predictors and their interactions was assessed using a type III Wald Chi-squared test (see Appendix B for model summary and Wald Chi-squared significance tests).

Results

Model results showed main effects of condition ($\chi^2(2) = 28.01$, $p < 0.001$) and block ($\chi^2(1) = 5.75$, $p < 0.05$). There was also an interaction of condition and group ($\chi^2(2) = 6.08$, $p < 0.05$). Overall, participants showed a strong priming effect (see Figure 3.1) for identity primes containing ambiguous /θ/=[θ/s] (b = -0.40, SE = 0.08, $p < 0.001$) and a weaker but significant priming effect for related (/s/-nonword) primes (b = -0.24, SE = 0.07, $p < 0.001$). Participants also showed faster responses in the second block of the experiment (b = -0.18, SE = 0.08, $p < 0.05$). Participants in the exposure group showed a numerically larger identity priming effect and smaller related priming effects (see Figure 3.1) but neither of these reached significance, despite the fact that the overall contribution of the group by block interaction term in the model was significant (as measured by a Wald Chi-squared test).

These results suggest that identity and related priming were observed for both groups of listeners. Given the apparent differences in the size of identity vs. related effects across control and training groups, the same model was fitted with this predictor releveled, such that the new reference level was set to be identity prime trials (rather than unrelated ones). Results show less priming for related primes compared to identity primes (b = 0.16, SE = 0.08, $p < 0.05$), but the interaction of condition and group suggest that this effect was stronger for participants with prior accent exposure (b = 0.26, SE = 0.11, $p < 0.05$). Given the numerically close identity priming effect across groups (Figure 3.1), this difference suggests that /s/ words served as poorer exemplars of /θ/ for trained listeners.

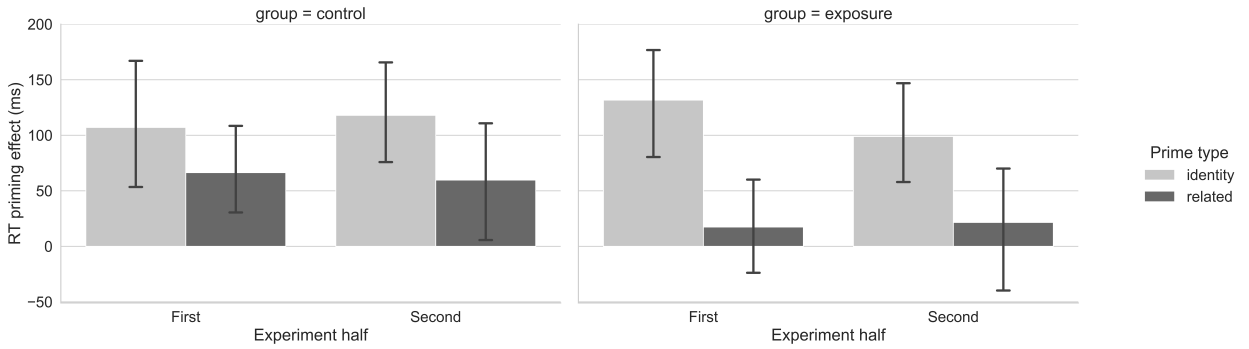


Figure 3.1: Experiment 3 cross-modal priming task results: Priming effect for identity and related trials, for groups exposed to critical /θ/ = [θ/s] words (experimental group) vs. replacement filler words (controls). Priming effects were calculated by subtracting RT for identity trials (e.g., [θ/*s*]*erapy* + <therapy>) and related trials (e.g., *serapy* + <therapy>) from trials with unrelated primes (e.g., *banana* + <therapy>. Error bars represent bootstrapped 95% confidence intervals.

Discussion

Results of Experiment 3 indicate that both control and experimental groups showed strong identity priming for ambiguous /θ/ = [θ/s] items, and a weaker but significant related priming effect for /θ/ = [s] primes. This indicates that experiments dependent on measures such as response latencies can be successfully conducted via web-based platforms, despite the additional variance introduced by non-laboratory testing conditions.

No effect of training was observed with identity primes, so we cannot conclude that exposure to such critical accented items caused significant changes in online lexical processing of the same accent. This null effect may be due to several factors. First, controls showed strong identity priming effects even in the first half of the experiment, and these stayed relatively stable through the second half of the experiment. This suggests that phonetically ambiguous /θ/ tokens in this experimental context did not pose a strong processing impediment, preventing possible differences between control and exposure groups from emerging due to ceiling effects. This observation is consistent with results by Marslen-Wilson et al. (1996) which showed that ambiguous primes such as *[t/d]ask* could prime *task* just as well as true identity primes. The exposure group, however, did show numerically stronger identity priming, and significantly weaker related priming compared to identity priming. Such results suggest a relatively specific processing strategy for trained listeners, who became less tolerant of atypical /θ/ pronunciations that did not match the exposure accent.

Even when /θ/ when was replaced with /s/ (in the related priming condition), participants still showed a consistent priming effect, indicating that the perceptual system maintained lexical activation of /θ/ items even when the auditory input was unambiguous. This result is consistent with the relative proximity of /θ/ and /s/ in perceptual space (see Figure

2.2).

3.4 Experiment 4

Hypotheses

In the following experiment, the effect of perceptual and acoustic similarity was further tested: listeners were trained on the same /θ/ = [θ/s] pronunciation but this time tested on a novel accent /θ/ = [θ/ʃ], produced by the same speaker. Given that /θ/ and /ʃ/ are less perceptually confusable than /θ/ and /s/ (Figure 2.2), this pronunciation ought to result in a weaker priming effect for controls with no prior exposure to an atypical /θ/ pronunciation. However, participants previously exposed to the /θ/ = [θ/s] accent should still see a training benefit, as measured by stronger identity priming for ambiguous /θ/ = [θ/ʃ] primes, assuming that changes in category structure facilitate lexical processing. Given the trend for weaker related (/θ/ = [s]) priming in the exposure group, we may also observe a similar pattern for related (/θ/ = [ʃ]) priming in Experiment 4.

Experimental platform

As in Experiment 3, participants were recruited via the Prolific online platform.

Participants

79 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 11 participants, leaving a total of 68 whose data were retained for analysis. As in Experiment 3, all participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. An additional 21 subjects were recruited on Prolific under the same selection criteria to participate in a pretest, which was used to select critical /θ/ tokens to be used in the main experiment. Participants were relatively evenly split on gender (F = 34, M = 32), with a small number selecting the 'other' response option (N = 2). Participants self-reported the following ethnicities: white (N = 52); black (N = 7), hispanic (N =5), asian (N = 3), and other (N = 1). They spanned the following age groups: 18-20 (N = 5), 21-25 (N = 15), 26-30 (N = 13), 31-35 (N = 15), 36-40 (N =6), 41-45 (N = 3), 46-50 (N = 3), 51+ (N = 8). They reported moderate experience with accented English on a seven-point scale (M = 4.8, SD = 1.8, range = 1-7).

Stimuli

The same training materials used in Experiments 1-3 were used in this study. The same speaker used to record training materials also produced all test materials. Materials in

the test phase consisted of the same 180 written targets (90 words + 90 nonwords) used in Experiment 3, paired with 180 auditory primes (90 words + 90 nonwords). Auditory primes were identical to those used in Experiment 3, with the exception that critical identity trials had a different artificial accent /θ/ = [θ/ʃ] from that used in the training phase (e.g., */θ/ʃ/erapy* + <therapy>), and critical related primes involved a substituted [ʃ] rather than [s] (e.g., *sherapy* + <therapy>). As in Experiment 3, critical auditory critical primes with ambiguous /θ/ (/θ/ = [θ/ʃ]) were created using the same morphing procedure and a separate norming task to determine selection of the most ambiguous token.

Procedure

Pretest

To select the most ambiguous critical /θ/ items for the exposure task, the 60 /θ/ - /ʃ/ continua were subjected to a pretest by a separate listener group, following the procedure used in Experiment 3. Of the 21 participants initially recruited via Prolific, 3 were removed due to failure to complete the task, and 1 due to inability or unwillingness to perceive the difference between items (<30% mean difference in responses between continuum endpoints). The selected tokens for each of the 60 continua were similar to the results from Experiment 3: the average step number of the ambiguous token was near the middle of the seven-step continuum (5.0), and the mean percentage of /θ/-word responses at this step was 62%.

Screening

Participants had to pass the same headphone check used in Experiment 3 prior to participation in the experimental task. Questionnaire results suggest that screening was again highly effective, with the majority of participants self-reporting use of high-quality headphones (N = 34) or ear-buds (N = 30), and only a small number passing the task using an external speaker (N = 4).

Exposure (lexical decision)

Participants completed the same training task used in Experiment 3, and were again randomly assigned to a training or control group.

Test (cross-modal priming)

Following the lexical decision task, participants completed the same cross-modal priming task used in Experiment 3, with the exception that identity primes were replaced with a different accent /θ/ = [θ/ʃ], and related primes involved substitution of /θ/ with /ʃ/.

Questionnaire

After completing the categorization task, participants were asked to complete the same questionnaire used in previous experiments.

Data preparation

Following the same procedure used in Experiment 3, data were first preprocessed by excluding participants who did not meet experimental criteria: 6 participants were removed for failing to complete the questionnaire, 3 for self-reporting as non-native speakers of English, 1 for failing to meet the 70% accuracy threshold in the test (cross-modal priming) task, and 1 for responding too quickly (RT < 200 ms) on over 20% of trials. Resulting data were subsequently cleaned to remove trials responses that were too fast (RT <200 ms), resulting in the removal of 216 trials (1.76%) of data. Finally, incorrect responses were removed prior to analysis (513 trials or 4.3% of data). This left a total of 11,511 observations (from 68 participants) that were retained for the analysis.

Analysis

Analyses were identical to those in Experiment 3. See Appendix B for model summary and Wald Chi-squared significance tests).

Results

Results showed significant main effects of condition ($\chi^2(2) = 22.26$, $p < 0.001$) and block ($\chi^2(1) = 4.29$, $p < 0.05$). There was also a significant interaction of condition and block ($\chi^2(2) = 13.57$, $p < 0.01$). No other terms reached significance.

Overall, participants responded faster to /θ/-target words when these were preceded by a ambiguous /θ/ = [θ/ʃ] "identity" primes, e.g., *[θ/ʃ]erapy* + <therapy> (b = -0.25, SE = 0.06, $p < 0.001$). Controls showed no priming effect with *sherapy* + <therapy> trials in the first block of the experiment (b = -0.03, SE = 0.06, p = 0.599). By contrast, the training group did show a significant priming effect here, as shown by a condition by group interaction (b = -0.18, SE = 0.09, $p < 0.05$). Participants responded faster overall in the second block of the experiment (b = -0.14, SE = 0.07, $p < 0.05$), and the interaction of block and condition indicate significantly increased identity priming (b = -0.28, SE = 0.08, $p < 0.001$) and related priming effects (b = -0.23, SE = 0.08, $p < 0.01$) for the control group in the second half of the experiment, which was not observed for the exposure group.

As in Experiment 3, to assess the difference between identity and related priming, the condition predictor was releveled, with the new reference level set to be identity prime trials rather than unrelated ones. Model results show that related primes yielded weaker priming effects than identity primes (b = 0.22, SE = 0.06, $p < 0.001$). Although this difference was attenuated for participants with prior accent exposure (suggesting comparatively stronger

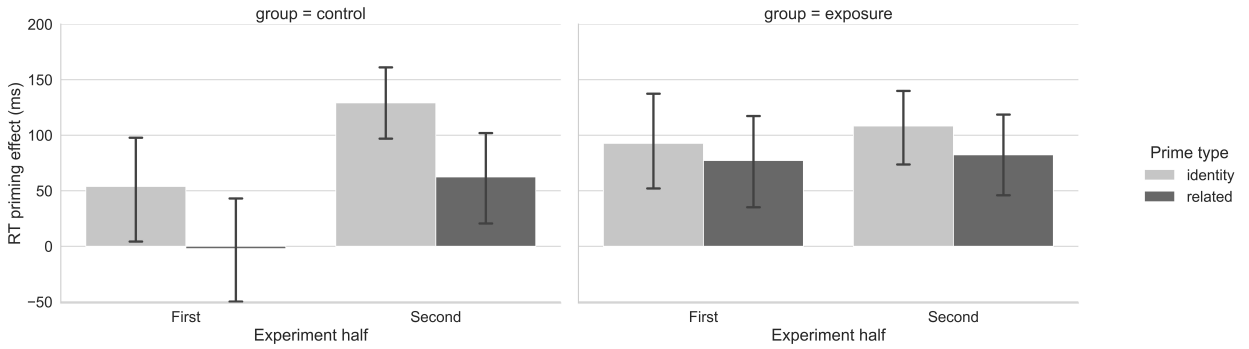


Figure 3.2: Experiment 4 cross-modal priming task results: Priming effect for identity and related trials, for groups exposed to critical /θ/ = [θ/s] words (experimental group) vs. replacement filler words (controls). Priming effects were calculated by subtracting mean RT for identity trials (e.g., [θ/ʃ/*erapy* + <therapy>) and related trials (e.g., *sherapy* + <therapy>) from trials with unrelated primes (e.g., *banana* + <therapy>. Error bars represent bootstrapped 95% confidence intervals.

related priming), this interaction of prime type and experimental group did not reach significance (b = -0.11, SE = 0.08, p = 0.20).

Exp.4 Discussion

These results provide tentative support for a faciliative effect of training on lexical processing. First, analysis of priming effects across controls and trained listeners showed significantly larger related priming for participants with prior accent exposure, in the first half of the experiment. There was also a numerically larger but non-significant effect of identity priming in the first experiment half. Results also showed that control participants saw perceptual learning over the course of the experiment, as indicated by significantly larger identity and related priming effects in the second experiment half. The experimental group saw no such learning effect — priming effects in both conditions were relatively stable across the two halves of the experiment. This pattern has several possible interpretations. The first is that prior accent exposure prevented learning of the new accent for listeners in the experimental group. This seems unlikely for several reasons. First, prior work has shown that listeners can rapidly adapt to changes in accent. For instance, a series of experiments by Kraljic and Samuel (2005, 2007) has shown that the recalibration effect can be "unlearned" by exposing listeners to good tokens of the ambiguous target sound. This is consistent with the observation that the recalibration effect tends to fade over the course of the categorization task (Charoy, 2021; Liu & Jaeger, 2019). The second interpretation is that learning from the exposure task could have immediately transferred for the trained participants, meaning their performance would have already been near-ceiling in the first block of the experiment. This would have left little room for improvement over the course of the task. Meanwhile,

by the second trial block controls already showed priming effects comparable to the training group. This interpretation would be consistent with previous recalibration work showing that learning can occur after exposure to just 10 critical words (Kraljic & Samuel, 2007). The significant increase in related priming is also consistent with the finding of Witteman et al. (2013) that priming for strongly-accented items (involving a categorical vowel substitution) was not seen initially but emerged in the second half of the cross-modal priming tasks. Finally, comparison of the identity priming effects across Experiments 3 and 4 show little difference in both the size and stability of priming across the task.

Overall, these results suggest that there was some transfer of perceptual learning due to training, consistent with the observation that exposure to an ambiguous /θ/ = [θ/s] pronunciation leads to categorization shifts for both /θ/-/s/ and /θ/-/ʃ/ continua. Given the weak evidence of a training benefit for related priming in Experiment 3, it is surprising to see one in this experiment. In fact, the size of related priming for trained listeners in the first half of the experiment was comparable to the size of identity priming for controls in that same block in Experiment 3, suggesting that training led to some degree of category expansion — even unambiguous /θ/ = [ʃ] words saw a relatively strong priming effect for listeners with prior exposure to the atypical /θ/ = [θ/s] pronunciation. The reason for this discrepancy in the related priming effect across experiments could be due to the nature of the exposure task. Both control and training groups heard critical /s/ words during exposure, meaning both groups learned what a typical /s/ production for this speaker ought to sound like. The training group also heard examples of ambiguous /θ/ = [θ/s], so they were familiarized with what a typical /θ/ production for this speaker ought to sound like. For this group, an /s/ thus served as a poorer example of /θ/ than it did for controls. Meanwhile, in Experiment 3, neither the training nor control group had heard tokens containing /ʃ/ during the exposure phase. It appears that exposure to ambiguous /θ/ thus resulting in this group showing more tolerance for unambiguous /θ/ = [θ/ʃ] productions during lexical access. Interestingly, it appears that the size of this priming effect for trained listeners was on par with identity primes — both priming effects were of comparable magnitude.

3.5 General discussion

Together, this set of experiments suggests that there is a relationship between changes in category structure and accompanying changes in lexical processing. Interestingly, this relationship emerges primarily in the analysis of between-group differences in the amount of related priming seen across experiments and across trial blocks within each experiment. In Experiment 3, results showed that both groups saw some priming in *serapy* + <therapy> trials. While this priming effect was weaker than in identity *[θ/s]erapy* + <therapy> trials for either group, listeners with prior accent exposure showed a larger difference between identity and related prime trials, suggesting that training led /s/ to be processed as a poorer example of /θ/. In Experiment 4, controls saw no priming for *sherapy* + <therapy> trials until the second half of the experiment, whereas the exposure group saw robust related

priming throughout the task.

Across experiments, there was no reliable difference between identity priming across groups, although participants with prior accent experience did show numerically stronger identity priming for both tasks. This result may be due to the fact that these "ambiguous" [θ/s] and [θ/ʃ] realizations of /θ/ did not pose serious impediments to lexical processing even for listeners who had no prior experience with this accent. Thus, ceiling effects for such primes could have prevented possible effects of training from emerging. This result lines up with previous work showing that a weak non-native accent does not always pose major processing difficulty; robust lexical activation of target words can be seen even with listeners who have limited prior experience with a given accent (Witteman et al., 2013).

Taken together, this set of result offers tentative support for a limited form of category expansion as a mechanism for perceptual adaptation to an unfamiliar accent. The different pattern of results seen for related priming across experiments and across experimental groups suggests, consistent with the results of Experiments 1 and 2, that perceptual similarity between training and test items played an important role in affecting the degree of lexical activation of the /θ/ targets. Overall, it appears that [s] served as a poorer example of /θ/ for participants with prior accent exposure, as compared to listeners with no prior accent exposure. This suggests that training resulted in changes to category representations for /θ/ that made listeners *less* tolerant of atypical pronounciations of this sound that differed from what they had previously encountered. Meanwhile, in Experiment 4, trained listeners seemed to process /ʃ/ as a relatively good instance of /θ/. In fact, the size of related priming for this group was nearly identical to that observed for identity priming, whereas controls showed no evidence that /ʃ/ words primed /θ/ targets at all (at least in the first half of the experiment). However, controls did adapt over the course of the experiment, showing significantly stronger identity and related priming in the second half of the experiment, whereas trained listeners showed no such effect, with related and identity priming effects remaining stable throughout the task.

Such results offer tentative support for a relation between changes to category structure and changes to lexical processing. Listeners with prior accent exposure showed some evidence of increased lexical activation when processing accented words produced with a novel accent that was similar to the exposure accent. However, as in Experiments 1 and 2, the set of experiments in this chapter show that the degree to which listeners are able to adapt to such pronunciations may be mediated by perceived similarity to the training accent—the size of the priming effect differed for /s/ vs. /ʃ/ primes. For controls, this difference is explicable in terms of perceptual similarity: /s/ is more perceptually confusable with /θ/ than /ʃ/ is (as previously visualized in Figure 2.2 in Chapter 2). It is thus reasonable to see related priming of /θ/ targets with /s/-word primes but not /ʃ/-word primes. For listeners with previous accent exposure, things are less clear. Given the shifts observed for both /θ/-/s/ and /θ/-/ʃ/ continua in Experiments 1 and 2, it is not clear why the degree of related priming (compared to identity priming) appears to differ between the two sounds. Possibly, this occurred because /ʃ/ was a perceptually closer match than /s/ to the training accent. Still, these results are interesting given that categorization results for both /θ/-/s/

and /θ/-/ʃ/ continua (Experiments 1 and 2) showed no category shift at all at the /s/ and /ʃ/ endpoints — both controls and trained listeners unambiguously processed these tokens as /s/ or /ʃ/.

The main results from this set of experiments aligns with prior work demonstrating a relationship between phonetic category structure and lexical processing. For instance, Xie et al. (2017) found that listeners exposed to an ambiguous word-final /d/ pronunciation in Mandarin-accented English were also more likely to process this sound as /d/ in a cross-modal priming task where critical primes involved ambiguous /d/-words that had /t/-word minimal pairs (e.g., auditory *see[d/t]* primed visual <seed>). However, the same devoiced /d/-final primes also continued to prime /t/-final visual targets, indicating that perceptual learning for this sound did not eliminate lexical competition for similar-sounding words. However, Xie et al. (2017) did not expose listeners to examples of /t/ in the training phase, meaning listeners did not have evidence for what a typical realization of /t/ sounded like for this speaker. By contrast, in the current set of studies, listeners in the training group heard both ambiguous /θ/ = [θ/s] word and unambiguous /s/ = [s] words. While prior exposure to both sounds did not eliminate /s/ word priming of visual /θ/ targets entirely, it did result in these listeners showing weaker related priming relative to identity priming, as compared to controls. However, listeners did not hear any instances of /ʃ/ in the training task. Given this fact and knowing that this speaker produced idiosyncratic /θ/, they may have been more flexible in allowing [ʃ] to map on to /θ/.

These results have several important implications for the question of the mechanisms listeners use to adapt to an unfamiliar accent. First, the finding that trained listeners, but not controls, showed priming of /θ/ words following exposure to auditory /ʃ/ minimal non-words suggests that listeners relied on some form of category expansion to adapt to the unfamiliar accent. Second, the fact that trained listeners also showed poorer related priming with /s/ minimal non-words indicates that there were limits on these listeners' tolerance for atypical realizations of the target sound. Overall, this suggests that listeners utilize a somewhat coarse-grained adaptation strategy for an unfamiliar accent. Perceptual learning involves adjustments to phonetic categories that are not limited to the phonetic patterns in the initial exposure period, but can also facilitate subsequent processing of similar but distinct pronunciations. Given the increased variability of accented speech (Wade et al., 2007), maintaining this kind of relatively general learning strategy may be important for generalizing learning to novel speakers.

Chapter 4

Investigating generalization of phonetic learning to new speakers

4.1 Introduction

This chapter investigates the factors affecting generalization of an unfamiliar accent to novel speakers. Chapter 2 results suggested that listeners use a non-uniform category expansion mechanism to adapt to a speaker with an unfamiliar pronunciation. Moreover, findings from Chapter 3 suggest that there is a link between category changes and processing — listeners with prior exposure to the accent showed changes in lexical processing for critical words produced with a novel accent similar to the training pronunciation. The experiments in the current chapter build on these findings by investigating whether those same mechanisms constrain the generalization of learning to novel speakers with a similar pronunciation.

Previous literature has suggested that this type of perceptual learning (lexically-guided phonetic recalibration) is often speaker-specific — it does not tend to generalize. However, results from Chapter 2 suggest that there is some listener tolerance for phonetic mismatch between exposure and test contexts — listeners generalize learning to a novel phonetic contrast containing the trained sound category. Chapter 3 provides complementary evidence, showing that even unambiguous /ʃ/ (e.g., *sherapy*) could prime /θ/ words (<therapy>) for listeners with prior exposure to an atypical pronunciation of /θ/ as [θ/s]. Such results suggest that lexically-guided recalibration of phonetic categories is sufficiently robust to phonetic variation to facilitate speaker-independent adaptation to a given accent, as has been found in related literature on accent accommodation (Bradlow & Bent, 2008; Sidaras et al., 2009). Crucially, given the expected phonetic variation across speakers (even when they share a language background), listeners must be able to abstract over such differences in order to successfully learn the accent. The literature shows that accented speech in particular tends to be more variable (Wade et al., 2007), but there are nonetheless systematic regularities within a given accent due to the influence of the L1 sound system. So, maintaining the right balance between perceptual flexibility and sensitivity to the systematic phonetic patterns

within a given accent may be crucial for speaker-independent adaptation to a given accent.

4.2 Background

Generalization in phonetic recalibration studies

The long-lasting nature of perceptual learning for speech, as Samuel and Kraljic (2009) note, raises the question of whether such learning is speaker-specific or not. If speaker-general, such learning could potentially inhibit accurate perception of speakers who do not share the pronunciation. The literature shows that the answer to this question is complicated, depending on the sound categories involved in training and test contexts, acoustic similarity between speakers, and the range of stimuli listeners are exposed to in training and test contexts. For instance, Kraljic and Samuel (2005) showed that perceptual learning-induced changes following exposure to ambiguous /s/ or /ʃ/ (pronounced as [s/ʃ]) persisted 25 minutes after exposure, even when listeners heard many canonical pronunciations (by a different speaker) in the intervening period. However, when listeners heard canonical forms produced by the same (idiosyncratic) speaker as in the original exposure, perceptual learning-induced changes in categorization were attenuated. The fact that 'unlearning' was only observed when listeners heard good tokens produced by the original speaker suggests speaker-specificity. Eisner and McQueen (2005) also found that learning of such atypical fricative pronunciations seemed to be speaker-specific. They found that generalization to different speakers was only possible provided that the fricatives used in a novel speakers' test continua were spliced in from the exposure speaker's productions. No transfer of learning was observed when the novel speakers' test continua used ambiguous fricatives created from their own productions.

Other research suggests that the specificity of phonetic recalibration may depend on the sound categories involved. Kraljic and Samuel (2006) found that learning of ambiguous stop consonants was much less constrained than previously reported results for fricatives. They exposed listeners to an ambiguous stop consonant [d/t] and found that learning generalized to both a novel speaker and a novel stop contrast, inducing a categorization shift for [b]-[p] continua. The authors suggest this difference is due to the fact that fricatives contain spectral information that can cue speaker identity, whereas stops do not — the implementation of stop voicing does not differ systematically across speakers. Kraljic and Samuel (2005) found that phonetic recalibraton following exposure to an ambiguous fricative pronunciation can transfer across speakers under certain conditions. In their study, they showed that learning for an ambiguous [s/ʃ] pronunciation transferred from a female training voice to male test voice, but not vice versa. Acoustic analysis of fricative spectral means in each set of training and test materials suggest that listeners were sensitive to acoustic similarity of fricatives between training and test: generalization occurred with more acoustically similar sets of training vs. test tokens (female training - male test) but not with more distinct ones (male training - female test).

Other literature on phonetic recalibration has also found that acoustic similarity between familiar (training) and novel (test) contexts facilitates cross-speaker generalization. Reinisch and Holt (2014) found that perceptual learning for an ambiguous fricative pronunciation could transfer to new speakers under certain conditions. They exposed American English listeners to a Dutch-accented female speaker of English with an ambiguous fricative production phonetically between [f] and [s]. They initially found that listeners were able to generalize learning to a novel female speaker, but not to a novel male speaker. However, a followup experiment showed that cross-gender generalization was possible when the familiar and novel speakers' test continua spanned a similar range of phonetic space.

Generalization of perceptual learning in natural accent accommodation studies

The literature on natural-accent accommodation, although utilizing different methodologies and types of speech exposure, has yielded similar results. Overall, patterns of generalization of learning for accented speech tends to to be mixed—learning does not consistently transfer to novel speakers. As with recalibration studies, the evidence here points to acoustic similarity as being an important factor for successful generalization.

Single-speaker exposure is often sufficient for listeners to show improved perception of a novel speaker. Xie, Weatherholtz, et al. (2018) found that a brief period of exposure to Mandarin-accented English led to both improved accuracy and faster response times in a cross-modal word recognition task, replicating a study by Clarke and Garrett (2004). Xie, Weatherholtz, et al. (2018) also demonstrated that such learning could transfer to a novel speaker sharing the same (Mandarin) accent. This adds to a set of studies demonstrating that adaptation to naturally-accented speech can transfer with minimal exposure to just a single speaker of that accent. For instance, Weil (2001) showed that listeners exposed to a single Marathi-accented speaker of English showed improved transcription accuracy for sentences produced by a novel Marathi-accented speaker. Witteman et al. (2013) showed that minimal exposure to German-accented Dutch (just 12 accented items) resulted in improvements in lexical processing of this accent, and that learning transferred to a novel German-accented Dutch speaker.

However, Bradlow and Bent (2008) found that single-speaker exposure was insufficient for speaker-independent accent accommodation. They exposed American English listeners to Mandarin Chinese-accented English spoken with a strong accent, with one listener group trained on many different speakers (high variability condition) whereas the other was trained on just one speaker. Results of a post-training test phase showed that both the high-variability and low-variability group showed approximately the same increase in transcription accuracy for a familiar Chinese-accented speaker compared to baseline (a control group with no prior accent exposure). However, performance for listeners in the low-variability group was at baseline when they were tested on a new speaker, whereas improvements transferred for the high-variability group. Thus, listeners in the multiple-talker condition

were able to achieve speaker-independent perceptual adaptation, while those in the single-talker condition only showed perceptual learning for the exposure speaker—learning did not transfer. Interestingly, the two groups did not differ when tested on a Slovakian accent which neither had been exposed to in training. This suggests that perceptual learning for accented speech involves tuning in to shared features across a group of speakers, rather than a simple increase in perceptual flexibility that may promote recognition of any accented speaker. These results suggest that high-variability training protocols may be important for achieving robust learning of a given accent.

Other studies support this observation. For instance, Sidaras et al. (2009) utilized the same basic paradigm as Bradlow and Bent (2008). The authors exposed listeners to sentences produced by 6 speakers of Spanish-accented English, and then tested them on novel sentences and words produced by either 6 different speakers or the same group of speakers they heard in training. Results showed that perceptual learning transferred to the novel speakers at both the sentence and word level—performance for both groups was comparable and significantly better than a control group that had been exposed to just natively-accented English. Baese-Berk et al. (2013) investigated whether multi-speaker exposure could induce accent-independent perceptual learning that would generalize to novel accents that listeners had not been previously exposed to. Like Bradlow and Bent (2008), they included an exposure phase where listeners were asked to transcribe speech produced by multiple accented speakers. Unlike the former study, however Baese-Berk et al. (2013) included speakers from a wide range of different L1 backgrounds. They hypothesized that if exposed to a sufficient range of variability on relevant phonetic dimensions, listeners would be able to generalize perceptual learning to a speaker of a novel accent. The idea was that because L2 speakers of English tend to struggle with some of the same phonetic features of English (e.g., production of typologically uncommon segments, slower speech rate, failure to reduce vowels, etc.), the right kind of exposure would allow listeners to learn this "global" foreign accent and achieve "accent-independent" adaptation. They used the stimuli/procedures of Bradlow and Bent (2008) but divided their participants into 3 training groups, who were exposed to either 5 speakers with a Standard American English accent, 5 Mandarin-accented speakers, or 5 speakers with different accents (Thai, Korean, Romanian, Mandarin, and Hindi). Results showed that listeners in the multiple-accent group were able to generalize perceptual learning to a novel (Slovakian) accent, but listeners in the single-accent group did not see learning transfer—their performance did not differ significantly from the native-accent control group.

Given the increased variability of non-natively accented speech (Wade et al., 2007), it is not surprising that high-variability training appears to facilitate general accent adaptation. It is plausible that exposure to many speakers can provide listeners with an approximation of the range of possible pronunciations they may encounter across speakers of a given accent group, allowing them to abstract over the differences between speakers. However, more recent literature suggests that such abstraction may not be the key mechanism in producing speaker-independent adapation. Rather, exposure to more speakers makes it more likely that listeners will find an acoustic match between a novel and previously encountered speaker. Xie and Myers (2017), for instance, exposed listeners to Mandarin-accented English where

word-final /d/ sounds were produced such that they were ambiguous with /t/. In a series of experiments, American English listeners were exposed to this accent in either a multi-speaker training phase or one of two different single-speaker training phases. They were then tested on a cross-modal priming task to assess whether training affected processing of word-final /d/ and /t/ words. Results showed that listeners in the multi-speaker training condition showed generalization of learning to a novel speaker of Mandarin-accented English, but only one of the two single-speaker training groups saw transfer of learning. Subsequent analyses showed that, regardless of training type (single- or multi-speaker) transfer of learning was best predicted by acoustic similarity between training-test speaker pairs. A study by Alexander and Nygaard (2019) also illustrates how acoustic similarity seems to be an important factor in generalization, regardless of L1 accent background. First, they found that multi-speaker training generalized when the training and test accents matched, e.g., listeners trained on multiple Spanish-accented speakers showed improved perception of novel Spanish-accented speakers, but not Korean-accented ones. Results of multi-accent training were intermediate: benefits were shown for words produced by novel Spanish-accented speakers, but not Korean-accented ones. Followup analyses suggested that between-speaker acoustic similarity on particular segments (vowels) may have facilitated transfer, with more acoustically similar speakers showing more robust generalization, regardless of L1 background. Together, these results suggest that a high-variability exposure phase may promote generalization of learning, but that the reason for this benefit may not be that it allows for abstraction over speaker differences. Rather, multi-speaker exposure may just make it more likely for listeners to find a close acoustic match between familiar and novel speakers.

The current study

Given the mixed evidence for generalization of perceptual learning for accented speech, it is plausible that listeners may rely on distinct mechanisms for achieving speaker-specific vs. speaker-general accent adaption. Previous research has suggested that listeners may have access to different strategies during accent accommodation (Schmale et al., 2012; Weil, 2001). Choice of strategy may depend on multiple factors (type of exposure materials, duration of training, number of speakers and accents in training phase, etc.). For instance, a high-variability exposure task (e.g., Baese-Berk et al., 2013) may encourage listeners to adopt a more general learning strategy (category expansion), whereas with lower exposure variability (e.g., Bradlow and Bent, 2008) a more targeted mechanism may be preferable.

Existing research is unclear about how mechanisms may differ for speaker-specific learning vs. speaker-independent learning. To assess the nature of the mechanism underlying recalibration of phonetic categories, the experiments in Chapter 2 tested listeners on a series of sound contrasts contained the trained phoneme. Results showed that learning was not contrast-specific, but generalized to a novel contrast that was perceptually similar to the training accent. This result is consistent with a phonetic learning strategy that involves expansion of a perceptual category into neighboring phonetic space, following exposure to an atypical production of that category. This result suggests that the perceptual system main-

tains a certain amount of uncertainty or coarseness of learning. It is plausible that maintaining this kind of general learning strategy could help listeners achieve speaker-independent adaptation to a given accent — since there is natural variability across speakers, even those with the 'same' accent will exhibit some degree of phonetic difference in their productions. This may be especially important for adapting to a non-native accent, since such speech tends to be especially variable (Seibert, 2011; Wade et al., 2007).

However, existing literature does not provide a clear answer to the question of whether listeners rely on the same mechanisms for speaker-specific vs. speaker-independent phonetic learning. Previous recalibration literature exploring cross-speaker generalization of learning has only examined a single phonetic contrast involving the atypical target sound (Kraljic & Samuel, 2005, 2007; Reinisch & Holt, 2014). The patterns of generalization in those studies thus do not yield clear evidence about the learning mechanisms involved. The current study aims to test this question explicitly by testing cross-speaker generalization for multiple phonetic contrasts involving the trained target phoneme, which ought to yield more conclusive evidence on which mechanisms listeners use during generalization of learning.

Hypotheses

Assuming that listeners rely on a form of category expansion to recalibrate phonetic categories and that this same mechanism transfers to novel speakers, we would predict that training on an ambiguous accent /θ/ = [θ/s] would result in an increased proportion of /θ/ responses for both /θ/-/s/ and /θ/-/ʃ/ continua, whether these were produced by the same speaker or a novel speaker. However, given mixed results for cross-speaker generalization of ambiguous fricative pronunciations in earlier work and the importance of acoustic match in generalization of learning (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Reinisch & Holt, 2014), it is plausible that we may observe asymmetric patterns of generalization for novel male vs. female speakers (e.g., transfer of learning to the former but not the latter), given the systematic gender-based differences in spectral characteristics for fricatives.

For /θ/-/s/ continua, where the sounds in the contrast match the exposure accent (/θ/ = [θ/s]), prior literature would suggest limited cross-gender generalization under normal testing conditions (Reinisch & Holt, 2014), given the systematic differences in spectral characteristics of /s/ across male vs. female speakers (male speakers generally have lower frequency resonances due to a longer vocal tract (Mann & Repp, 1980; Strand & Johnson, 1996)). This suggests that, consistent with previous work, we may observe generalization to the more acoustically similar male speaker but not to the female speaker for this contrast.

For /θ/-/ʃ/ continua, given that existing literature on cross-speaker generalization of phonetic recalibration has only tested a single phonetic contrast following exposure, it is not clear what exact pattern of results we should expect. However, if listeners are using a non-uniform category expansion mechanism, and generalization is primarily due to raw acoustic similarity between training and test tokens, then we may again predict generalization to the novel male speaker (increased proportion of /θ/ responses). Transfer of learning to the female speaker could also occur, given that female speakers' fricative productions tend to

exhibit a systematically higher spectral mean. This means that this speaker's /θ/-/ʃ/ items may fall within the acoustic range of the male speaker's [θ/s] tokens. If this is the case, then we ought to observe transfer of learning here as well.

4.3 Experiment 5

Methods

Experimental platform

The experiments in this chapter were conducted using the same custom web-based program used in previous experiments 1-4, and all study procedures occurred through Prolific.

Participants

81 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 13 participants, leaving a total of 68 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants were relatively evenly divided on gender (F = 36, M = 29,), with a small number selecting the 'other' response option (N = 3). Participants self-reported the following ethnicities: White (N = 50), Hispanic (N = 8), Black (N = 7), Asian (N = 2), and 'other' (N = 1). They spanned the following age groups: 18-20 (N = 10), 21-25 (N = 18), 26-30 (N = 18), 31-35 (N = 9), 36-40 (N = 2), 41-45 (N = 2), 46-50 (N = 1), 51+ (N = 8). Overall, participants reported moderate experience with accented English on a seven-point scale (M = 4.9, SD = 1.7, range = 1-7).

Stimuli

The same training materials used in previous experiments were presented in this study, following the same experimental procedures. Test materials included the same /θ/-/s/ minimal-pair continua used in Experiment 1. Howeever, in this experiment the test continua were based on words produced by 3 different speakers: the male speaker who produced the training materials (age = 27), a novel male speaker (age = 30) and a novel female speaker (age = 25). All speakers were native speakers of American English, had grown up in California, and were living in California at the time of recording. Stimuli were recorded and processed using the same procedure as in Experiment 1.

Procedure

Screening

Participants had to pass the same headphone check used in previous experiments prior to participation in the experimental task. Questionnaire results suggest that screening was again highly effective, with the majority of participants self-reporting use of high-quality headphones (N = 37) or ear-buds (N = 29), and only a small number passing the task using an external speaker (N = 2).

Exposure (Lexical Decision)

Participants in the accent exposure group completed a lexical decision task, using the same materials and following the same procedure as in previous experiments.

Test (Categorization)

All participants completed a categorization task using the same 4 /θ/-/s/ minimal-pair continua presented in Experiment 1. Each continuum was presented twice for each speaker, resulting in a total of 168 tokens (3 speakers x 4 continua x 7 tokens x 2 repetitions per token). Tokens were presented in a different random order for each participant, following Reinisch and Holt (2014). The presentation of fewer tokens per speaker (compared to previous experiments where 4 repetitions per token were utilized) was chosen in order to minimize task fatigue and because recent research has suggested that the recalibration effect fades over the course of the testing phase (Charoy, 2021; Liu & Jaeger, 2018, 2019). Thus, presentation of fewer tokens ought to facilitate detection of the training effect. Aside from the differences noted here, testing procedures matched those utilized in Experiments 1 and 2.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 3 participants were removed for completing the task more than once, 6 were removed for failing to complete the questionnaire, 1 for failing to complete the training task, 1 for self-reporting as a non-native speaker of English, and 2 for failing to reliably perceive a difference between continuum endpoints ($< 50\%$ mean difference between continuum points 1 and 7). Resulting data were subsequently cleaned to remove responses that were too fast (RT <200 ms) or too slow (RT > 2500 ms), resulting in the removal of 581 trials (5.0%) of data.

Analysis

Data were analyzed via generalized linear mixed-effects regression modeling, using the lme4 package (Bates et al., 2014) in R (R Core Team, 2022). Maximal random effect structure was used for each model where this did not result in convergence issues, with random slopes

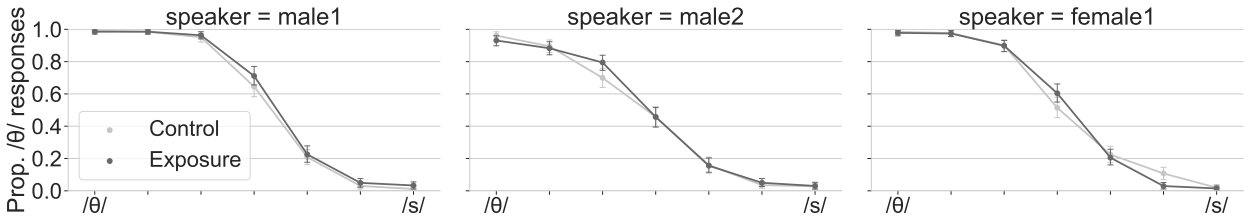


Figure 4.1: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [s] phonetic continua, by speaker and by exposure condition. No significant effect of training condition was found for any of the speakers, including the exposure speaker (male1).

fitted for all within-participant predictors (Barr, 2013). The significance of fixed effects and interactions between them was assessed using a Wald chi-squared test. Categorical variables were treatment-coded and included condition (training vs control, ref = control), /θ/ word position (word-initial vs. word-final, ref = word-final), and speaker (ref = male exposure speaker). A single numeric variable (continuum step, range = 1-7) was included as a centered numeric variable. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Results

Model results show significant main effects of continuum step ($\chi^2(1) = 197.26$, $p < 0.001$) and speaker ($\chi^2(2) = 25.86$, $p < 0.001$). There were also significant 2-way interactions of step by group ($\chi^2(1) = 4.90$, $p < 0.05$), step by speaker ($\chi^2(2) = 11.23$, $p < 0.01$), and word position by speaker ($\chi^2(2) = 65.32$, $p < 0.001$). Finally, there was one significant 3-way interaction of step, group, and word position ($\chi^2(1) = 4.96$, $p < 0.05$). However, the experimental group predictor was not significant on its own, nor was there an interaction of experimental group and speaker, indicating that the effect of training was not detected in this experiment.

As expected, participants classified fewer tokens as /θ/ as a function of continuum step (b = -2.27, SE = 0.16, z-value = -14.04, $p < 0.001$). They also showed less /θ/ responses overall for both the male generalization speaker (b = -0.88, SE = 0.21, z-value = -4.26, $p < 0.001$) and the female generalization speaker (b = -1.01, SE = 0.22, z-value = -4.62, $p < 0.001$). Significant interactions of the predictors step by speaker and word position by speaker suggest that overall listeners were sensitive to the difference between the three speakers, as reflected in different categorization functions (see Figure 4.1.

Discussion

These results are surprising given previous recalibration literature that has used a similar multi-speaker test paradigm to evaluate the cross-speaker generalization of perceptual learn-

ing for fricatives (Reinisch & Holt, 2014; Reinisch et al., 2013). Such previous work has suggested that intermixing tokens across speakers in this way may facilitate the perception of their common accent. In this experiment, by contrast, this method of presentation appeared to have the opposite effect, with the result that a training effect was not observed even for the familiar (exposure) speaker. The null effect observed here suggests that this testing paradigm may not be appropriate to evaluate phonetic recalibration with more than one novel speaker. Reinisch and Holt (2014) report significant differences in listeners' categorization responses for the familiar speaker based on whether she was tested in the context of a novel female or male generalization speaker. In their case, this change did not nullify the effect of training. However, it is possible that in the current experiment, including more than 2 speakers in the test phase led to more more response variability in listeners, preventing the detection of a possible effect.

4.4 Experiment 6

In order to test whether the null effect of training in the previous experiment was due to the novel test paradigm, a series of follow-up experiments were conducted. These analyses directly replicated the methods used in Chapter 1 — listeners in each experiment were all trained on the same speaker as in Chapter 1, and were only tested on a single speaker in the categorization task. This ought to avoid the possible confound of stimulus presentation in multi-speaker categorization tasks and ensure that listener categorization responses would not be affected by tokens produced by another speaker. Experiment 6a is a direct replication of Experiment 1, with both training and test materials (/θ/-/s/ continua) produced by the same speaker. Experiment 6b tests generalization of learning with the novel male speaker from Experiment 5, and Experiment 6c tests generalization of learning to the novel female speaker from Experiment 5. All experiments used the same custom web program as in previous experiments, with all study procedures occurring through the Prolific platform.

Experiment 6a - Male exposure speaker

Participants

87 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 15 participants, leaving a total of 72 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants were relatively evenly divided on gender (M = 35, F = 34), with a small number selecting the 'other' response option (N = 3). Participants self-reported the following ethnicities: White (N = 56); Black (N = 7), Hispanic (N = 5), Asian (N = 2), and other (N = 2). They spanned the following age groups: 18-20 (N = 3), 21-25 (N = 16), 26-30 (N = 16), 31-35 (N

= 10), 36-40 (N =8), 41-45 (N-6), 46-50 (N = 4), 51+ (N = 8). They reported moderate experience with accented English on a seven-point scale (M = 5.0, SD = 1.7, range = 1-7).

Stimuli

The same training used in Experiment 1 were used in this study. The same test materials used in Experiment 1 were also used, with the exception that participants were only tested on /θ/-/s/ continua (not /θ/-/f/). Otherwise, all experimental procedures matched those in Experiment 1.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 11 participants were removed for completing the task more than once, 2 were removed for failing to complete the questionnaire, and 2 for self-reporting as non-native speakers of English. Resulting data were subsequently cleaned to remove trials responses that were too fast (RT <200 ms) or too slow (RT > 2500 ms), resulting in the removal of 206 trials (2.47%) of data.

Analysis

Analyses were identical to those in Experiment 5, with the exception that experiment half was added as a categorical predictor (trial block 1 vs. block 2, ref = block 1). Recent studies have shown that the recalibration effect can diminish or disappear over the course of testing it (Charoy, 2021; Liu & Jaeger, 2018, 2019). Although not originally present in the analyses for Experiments 1 and 2, including this factor could avoid failure to detect a training effect due to a watering down of the effect size by trials occurring late in the task. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Experiment 6b - Male generalization speaker

Participants

88 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 16 participants, leaving a total of 69 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants were evenly divided on gender (M = 33, F = 30), with a small number selecting the 'other' response option (N = 3). Participants self-reported the following ethnicities: White (N = 46); Black (N = 8), Asian (N = 8), Hispanic (N = 5), and other (N = 2). They spanned the following age groups: 18-20 (N = 8), 21-25 (N = 15), 26-30 (N = 15), 31-35 (N

= 18), 36-40 (N =5), 41-45 (N = 2), 46-50 (N = 2), 51+ (N = 4). They reported moderate experience with accented English on a seven-point scale (M = 4.9, SD = 1.6, range = 1-7).

Stimuli

The same training materials used in Experiments 1-2 were used in this study, following the same experimental procedures. Test materials included the same 4 minimal-pair continua used in Experiment 6a, recorded by the novel male speaker in Experiment 5. Number of test stimuli and presentation procedure matched Experiment 6a.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 11 participants were removed for completing the task more than once, 3 were removed for failing to complete the questionnaire, 3 for self-reporting as non-native speakers of English, and 2 for failing to reach a 50% mean difference between continuum endpoints in the categorization task. Resulting data were subsequently cleaned to remove trials responses that were too fast (RT <200 ms) or too slow (RT > 2500 ms), resulting in the removal of 316 trials (3.97%) of data.

Analysis

Analysis was identical to that used in Experiment 6a. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Experiment 6c - Female generalization speaker

Participants

80 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 11 participants, leaving a total of 69 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants were primarily male (M = 42, F = 25), with a small number selecting the 'other' response option (N = 2). Participants self-reported the following ethnicities: White (N = 45), Asian (N = 9), Hispanic (N = 8), Black (N = 5), and other (N = 2). They spanned the following age groups: 18-20 (N = 10), 21-25 (N = 16), 26-30 (N = 10), 31-35 (N = 11), 36-40 (N =7), 41-45 (N = 2), 46-50 (N = 6), 51+ (N = 7). They reported moderate experience with accented English on a seven-point scale (M = 4.9, SD = 1.6, range = 2-7).

Stimuli

The same training materials used in Experiments 1-2 were used in this study, following the same experimental procedures. Test materials included the same 4 minimal-pair continua used in Experiment 6a-b, recorded by the novel female speaker speaker in Experiment 5. Number of test stimuli and presentation procedure matched that used in Experiments 6a-b.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 7 participants were removed for completing the task more than once, 2 were removed for failing to complete the questionnaire, and 2 for self-reporting as non-native speakers of English. Resulting data were subsequently cleaned to remove trials responses that were too fast (RT <200 ms) or too slow (RT > 2500 ms), resulting in the removal of 214 trials (2.72%) of data.

Analysis

Analysis was identical to that used in Experiment 6a-b. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Results

Exp. 6a - Male exposure speaker

Model results show significant main effects of continuum step ($\chi^2(1) = 165.54$, $p < 0.001$), experimental group ($\chi^2(1) = 6.43$, $p < 0.05$), and block ($\chi^2(1) = 6.32$, $p < 0.05$). There were also significant 2-way interactions of continuum step by group ($\chi^2(1) = 5.53$, $p < 0.05$), as well as continuum step by block ($\chi^2(1) = 4.21$, $p < 0.05$). Finally, there were also significant 3-way interactions of step, group, and block ($\chi^2(1) = 6.73$, $p < 0.01$), and step, word position, and block ($\chi^2(1) = 7.31$, $p < 0.01$).

The exposure group showed showed a significantly higher proportion (3.8%) of /θ/ responses (b = 0.86, SE = 0.34, z-value = 2.54, $p < 0.05$) compared to controls , with most of the shift apparent in just the middle steps of the continuum (see Figure 4.2), contrary to what was observed in Experiment 1, where the shift spanned the virtually all continuum steps. There was a trend for a decrease in the size of the training effect across experiment blocks (3.9% shift seen in the first half of trials, and a 3.6% shift in the second half), but this was not significant (b = -0.38, SE = 0.31, z-value = -1.23, p = 0.22). The model also predicted a lower likelihood of /θ/ responses closer to the [s] side of the continuum (b = -2.47, SE = 0.19, z-value = -12.87, $p < 0.001$), and a higher overall likelihood of /θ/ responses in the second block of the experiment (b = 0.55, SE = 0.22, z-value = 2.51, $p < 0.05$).

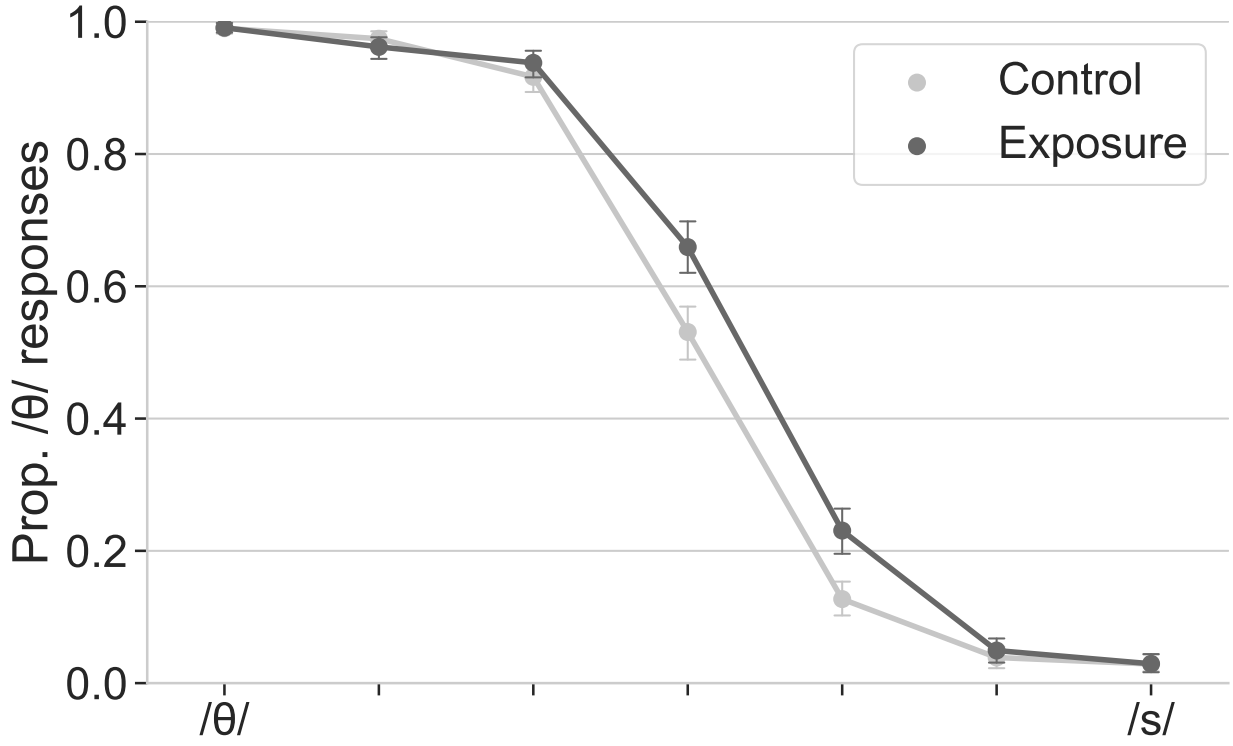


Figure 4.2: Male exposure speaker: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [s] phonetic continua, by exposure condition.

Exp. 6b - Male generalization speaker

Results showed main effects of continuum step ($\chi^2(1) = 173.19$, $p < 0.001$), word position ($\chi^2(1) = 18.80$, $p < 0.001$), and trial block ($\chi^2(1) = 8.77$, $p < 0.01$), with a marginal effect of experimental group ($\chi^2(1) = 3.28$, $p = 0.070$). There were also significant 2-way interactions of continuum step by word position ($\chi^2(1) = 7.41$, $p < 0.01$) as well as word position by block ($\chi^2(1) = 8.16$, $p < 0.01$). There was also a significant 3-way interaction of step, word position, and block ($\chi^2(1) = 11.14$, $p < 0.001$).

The exposure group showed a marginally significant (5.3%) increase in the proportion of /θ/ responses compared to controls (b = 0.57, SE= 0.31, z-value = 1.81, p = 0.070). Interestingly, the training effect spanned most of the continuum (see Figure 4.3), in contrast to what was seen with the exposure speaker, where it was localized to the ambiguous region of the continuum. This suggests that the novel male speaker had more acoustically similar /θ/ and /s/ productions to begin with, leaving room for a group difference to emerge even near continuum endpoints, as we saw in Experiment 1. There was a numeric difference in the size of the training effect across experiment blocks (a 6.9% shift seen in the first half of trials, and a 3.5% shift in the second half), but this difference was not significant (b = -0.30,

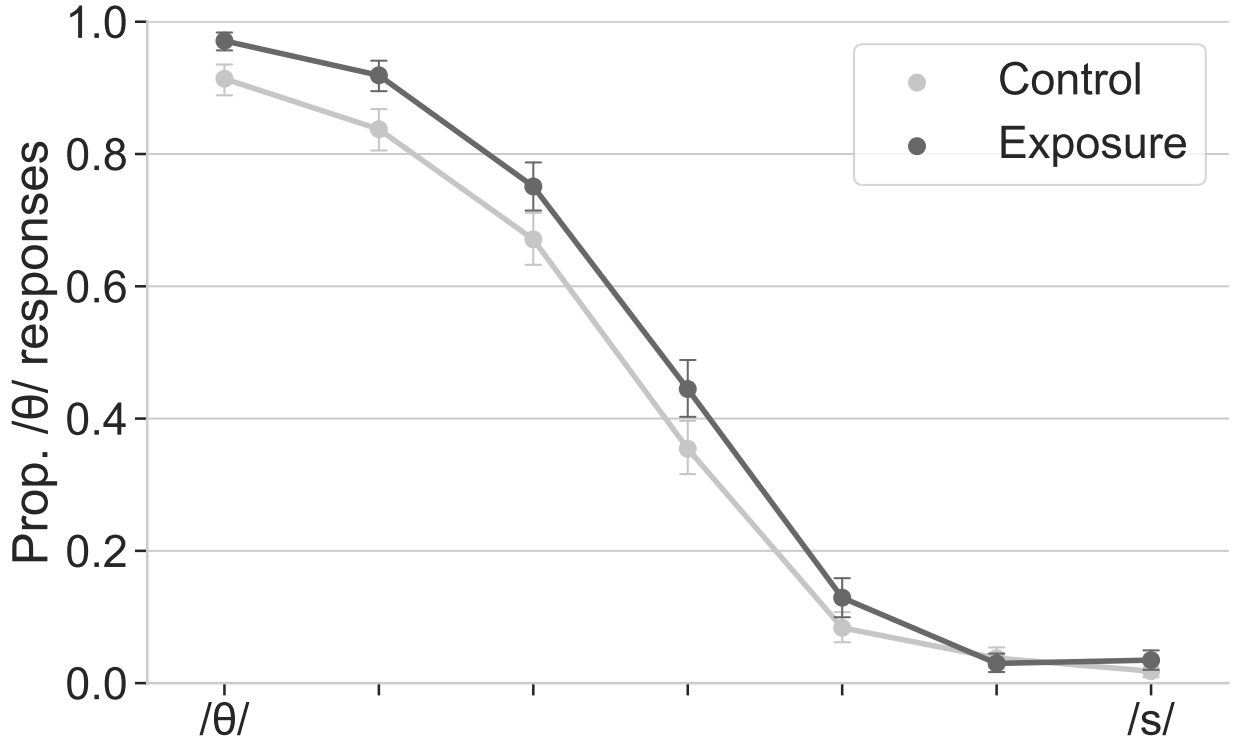


Figure 4.3: Male generalization speaker: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [s] phonetic continua, by exposure condition.

SE = 0.35, z-score = -0.85, p = 0.39).

Unsurprisingly, results again showed that proportion of /θ/ responses decreased with continuum step for tokens closer to the /s/ end-point (b = -1.70, SE = 0.13, z-value = -13.16, $p < 0.001$). They also showed a lower likelihood of /θ/ responses in word-initial /θ/ continua (b = -1.16, SE = 0.27, z-value = -4.34, $p < 0.001$). The model also predicted a higher likelihood of /θ/ responses in the second block of the experiment (b = 0.75, SE = 0.25, z-value = 2.96, $p < 0.01$). The interaction of step and word position suggest that the effect of step on decreasing the likelihood of /θ/ responses was weakened in word-initial /θ/ continua (b = 0.33, SE = 0.12, z-value = 2.72, $p < 0.01$), but a 3-way interaction of step, word position, and block suggest that this weakening effect was not as strong in the second block of the experiment (b = -0.66, SE = 0.20, z-value = -3.34, $p < 0.001$) The interaction of word position and block suggest that the predicted effect of word-initial position on decreasing the likelihood of a /θ/ response was weakened in the second block (b = 0.76, SE = 0.27, z-value = 2.86, $p < 0.01$).

Exp. 6c - Female generalization speaker

Results showed significant main effects of continuum step ($\chi^2(1) = 198.42$, $p < 0.001$) and of block ($\chi^2(1) = 18.91$, $p < 0.001$). There was also a significant 2-way interaction of group and block ($\chi^2(1) = 4.25$, $p < 0.05$).

The exposure group showed showed a small numeric shift (1.8%) in their /θ/-/s/ category boundary (see Figure 4.4, but this difference was not significant (b = 0.13, SE = 0.26, z-value = 0.48, p = 0.63). There was also a significant decrease in the size of the training effect across experiment blocks (b = -0.92, SE = 0.44, z-score = -2.06, $p < 0.05$), with a 2.4% shift seen in the first half of trials, and a 1.3% shift in the second half. As expected, the predicted likelihood of a /θ/ response decreased closer to the /s/ endpoint of minimal-pair continua (b = -2.28, SE = 0.16, z-value = -14.09, $p < 0.001$). There was also a higher predicted likelihood of /θ/ responses in the second block of the experiment (b = 1.36, SE = 0.31, z-value = 4.35, $p < 0.001$).

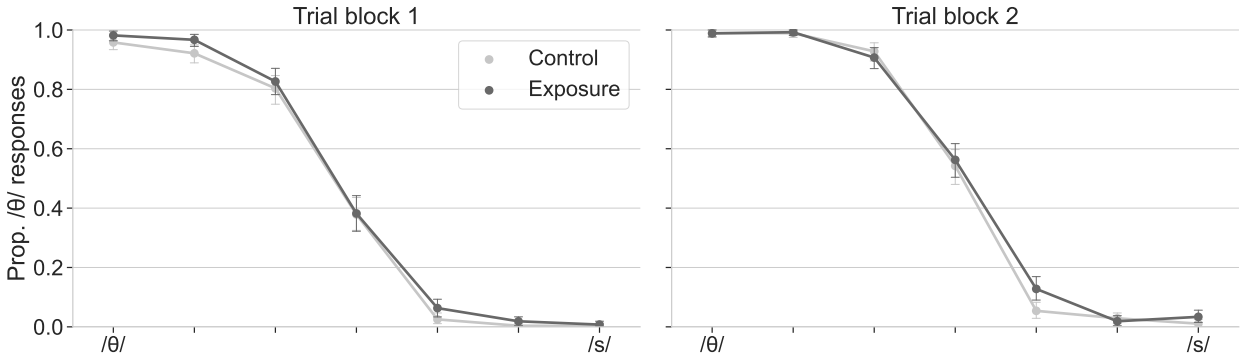


Figure 4.4: Female generalization speaker: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [s] phonetic continua, by exposure condition. A significantly small training effect was predicted in the second block of trials.

Experiment 6a-c discussion

Results of this set of experiments are consistent with prior work showing that generalization for phonetic recalibration of fricatives is relatively constrained. A significant boundary shift for the male exposure speaker successfully replicates results from Experiment 1, indicating an increase in proportion of /θ/ responses following exposure to an ambiguous /θ/ = [θ/s] pronunciation. Results for the novel male speaker showed a relatively large boundary shift consistent across continuum steps. comparable to that observed for the exposure speaker. However, model results indicate that the effect of training in this experiment only approached significance, so we cannot be confident that generalization of learning occurred here. Finally, the novel female speaker showed no evidence for generalization of learning, although there

was a small numeric increase in proportion of /θ/ responses for listeners in the training group.

These results are consistent with a number of previous phonetic recalibration studies showing that such learning for fricatives tends to resist cross-speaker generalization. For instance, Kraljic and Samuel (2005) showed that learning only transferred when exposure and test contexts were highly acoustically similar. Reinisch and Holt (2014) found a complementary result. In an initial set of experiments, they found generalization from a female exposure speaker to a novel female test speaker, but not to a novel male test speaker. A subsequent experiment presented a subset of the male speaker's original test continua, such that they sampled a perceptual space similar to that of the female speaker. Under these conditions, learning did transfer. The results of this set of experiments indicate that just as with previous studies involving different fricative contrasts, perceptual learning of a speaker with an ambiguous /θ/ pronunciation tended to be speaker-specific, albeit with some evidence for generalization for speakers of the same gender.

4.5 Experiment 7

The following set of experiments tests whether generalization of learning for an ambiguous /θ/ = [θ/s] pronunciation can transfer to both a new speaker and a new phonetic contrast /θ/-/ʃ/. Given that transfer of learning was observed within a single speaker in Chapter 2, it is possible that the same mechanism may faciliate cross-speaker generalization for the same contrast. Given gender-based spectral differences in fricative realization (Mann & Repp, 1980; Strand & Johnson, 1996), we might expect a different pattern of results here, since the female speaker's /ʃ/ may be acoustically closer to the male speaker's /s/, allowing transfer for this contrast even though no transfer was observed in /θ/-/ʃ/ categorization.

The set of experiments presented below each utilize the same training materials as in Experiments 6a-c, produced by the same speaker. The test phase, however, involves categorization of a /θ/-/ʃ/ continuum, as in Experiment 2. Experiment 7a is a direct replication of Experiment 2, with both training and test materials produced by the same speaker. Experiment 7b tests generalization of learning for the novel male speaker in Experiment 6b, and Experiment 7c tests generalization of learning to the novel female speaker from Experiment 6c. All experiments here are otherwise identical to Experiments 6a-c.

Experiment 7a - Male exposure speaker

Participants

78 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 6 participants, leaving a total of 72 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants

were relatively evenly divided on gender (F = 38, M = 32), with a small number selecting the 'other' response option (N = 2). Participants self-reported the following ethnicities: White (N = 58); Black (N = 5), other (N = 4), Hispanic (N = 3), and Asian (N = 2). They spanned the following age groups: 18-20 (N = 7), 21-25 (N = 14), 26-30 (N = 13), 31-35 (N = 7), 36-40 (N = 13), 41-45 (N = 2), 46-50 (N = 5), 51+ (N = 11). They reported moderate experience with accented English on a seven-point scale (M = 4.8, SD = 1.6, range = 1-7).

Stimuli

The same training materials and test materials (/θ/-/ʃ/ minimal-pair continua) used in Experiment 2 were used in this experiment, following the same experimental procedures.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 7 participants were removed for completing the task more than once, 3 were removed for failing to complete the questionnaire, 2 for self-reporting as non-native speakers of English, and 1 for responding too quickly (RT < 200 ms) on over 20% of trials. Resulting data were subsequently cleaned to remove responses that were too fast (RT < 200 ms) or too slow (RT > 2500 ms), resulting in the removal of 156 trials (1.91% of data).

Analysis

Analyses were identical to those used in Experiments 6a-c. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Experiment 7b - Male generalization speaker

Participants

79 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 8 participants, leaving a total of 71 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants were primarily male (M = 41, F = 28), with a small number selecting the 'other' response option (N = 2). Participants self-reported the following ethnicities: White (N = 44), Asian (N = 12), Black (N = 6), other (N = 5), and Hispanic (N = 4). They spanned the following age groups: 18-20 (N = 7), 21-25 (N = 23), 26-30 (N = 7), 31-35 (N = 12), 36-40 (N = 6), 41-45 (N = 3), 46-50 (N = 7), 51+ (N = 6). They reported moderate experience with accented English on a seven-point scale (M = 5.0, SD = 1.7, range = 1-7).

Stimuli

The same training materials used in Experiments 2 were used in this study, following the same experimental procedures. Test materials utilized the same /θ/-/ʃ/ minimal pair continua used in Experiment 2, but recorded by the male generalization speaker in Experiment 6b.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 2 participants were removed for completing the task more than once, 4 were removed for failing to complete the questionnaire, 2 for self-reporting as non-native speakers of English, and 1 for responding too quickly (RT $<$ 200 ms) on over 20% of trials. Resulting data were subsequently cleaned to remove responses that were too fast (RT $<$200 ms) or too slow (RT $>$ 2500 ms), resulting in the removal of 173 trials (2.16% of data).

Analysis

Analyses were identical to those used in Experiments 6a-c. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Experiment 7c - Female generalization speaker

Participants

79 participants were initially recruited on Prolific. Exclusion of participants who failed to meet experimental criteria resulted in the removal of 14 participants, leaving a total of 65 whose data were retained for analysis. All participants lived in the U.S., reported being native speakers of American English with normal speech and hearing, and gave informed consent prior to participating in the study. They were paid $15/h for their participation. Participants were primarily male (M = 40, F = 25). Participants self-reported the following ethnicities: White (N = 49), Hispanic (N = 8), Black (N = 4), Asian (N = 2), and other (N = 2). They spanned the following age groups: 18-20 (N = 10), 21-25 (N = 11), 26-30 (N = 13), 31-35 (N = 7), 36-40 (N = 9), 41-45 (N = 3), 46-50 (N = 4), 51+ (N = 8). They reported moderate experience with accented English on a seven-point scale (M = 4.7, SD = 1.4, range = 1-7).

Stimuli

The same training materials used in Experiment 2 were used in this study, following the same experimental procedures. Test materials utilized the same /θ/-/ʃ/ minimal pair continua used in Experiment 2, recorded by the female generalization speaker in Experiment 6c.

Data preparation

Prior to statistical analysis, data were first preprocessed by excluding participants who did not meet experimental criteria: 2 participants were removed for completing the task more than once, 4 were removed for failing to complete the questionnaire, 2 for self-reporting as non-native speakers of English, and 1 for responding too quickly (RT < 200 ms) on over 20% of trials. Resulting data were subsequently cleaned to remove responses that were too fast (RT <200 ms) or too slow (RT > 2500 ms), resulting in the removal of 163 trials (2.21% of data).

Analysis

Analyses were identical to those used in Experiments 6a-c. See Appendix C for a summary of model estimates and results of the Wald Chi-squared test.

Results

Exp. 7a results

Model results revealed main effects of continuum step ($\chi^2(1) = 177.03$, $p < 0.001$) and block ($\chi^2(1) = 31.10$, $p < 0.001$). There was also a significant 2-way interaction of word position and block ($\chi^2(1) = 8.77$, $p < 0.01$), and a 4-way interaction of step, group, word position, and block ($\chi^2(1) = 6.00$, $p < 0.05$). Crucially, the effect of experimental group failed to reach significance ($\chi^2(1) = 2.38$, $p = 0.12$), failing to replicate the finding of Experiment 2, where a significant effect of training was observed for categorization of [θ]-[ʃ] continua.

There was a trend for an increased proportion of /θ/ responses for the training group (see Figure 4.5), but this result was not significant (b = 0.58, SE = 0.38, z-value = 1.54, p = 0.12). There was a small numeric difference in the size of the training effect across experiment blocks (a 1.8% shift seen in the first half of trials, and a -1.4% shift in the second half), but model results showed no significant effect of block on the training effect ($\chi^2(1) = 0.26$, $p = 0.61$).

As expected, the model predicted a lower likelihood of /θ/ responses toward the /ʃ/ endpoint of the continuum (b = -2.61, SE = 0.20, z-value = -13.31, $p < 0.001$), and higher likelihood of a /θ/ response in the second block of trials (b = 1.80, SE = 0.32, z-value = 5.58, $p < 0.001$). The interaction of word position and block also predicts a lower likelihood of /θ/ responses for word-initial continua in the second block of trials (b = -1.17, SE = 0.40, z-value = -2.96, $p < 0.01$).

Exp. 7b results

Model results revealed main effects of continuum step ($\chi^2(1) = 127.52$, $p < 0.001$), experimental group ($\chi^2(1) = 6.78$, $p < 0.01$), word position ($\chi^2(1) = 63.24$, $p < 0.001$), and trial block ($\chi^2(1) = 22.06$, $p < 0.001$).

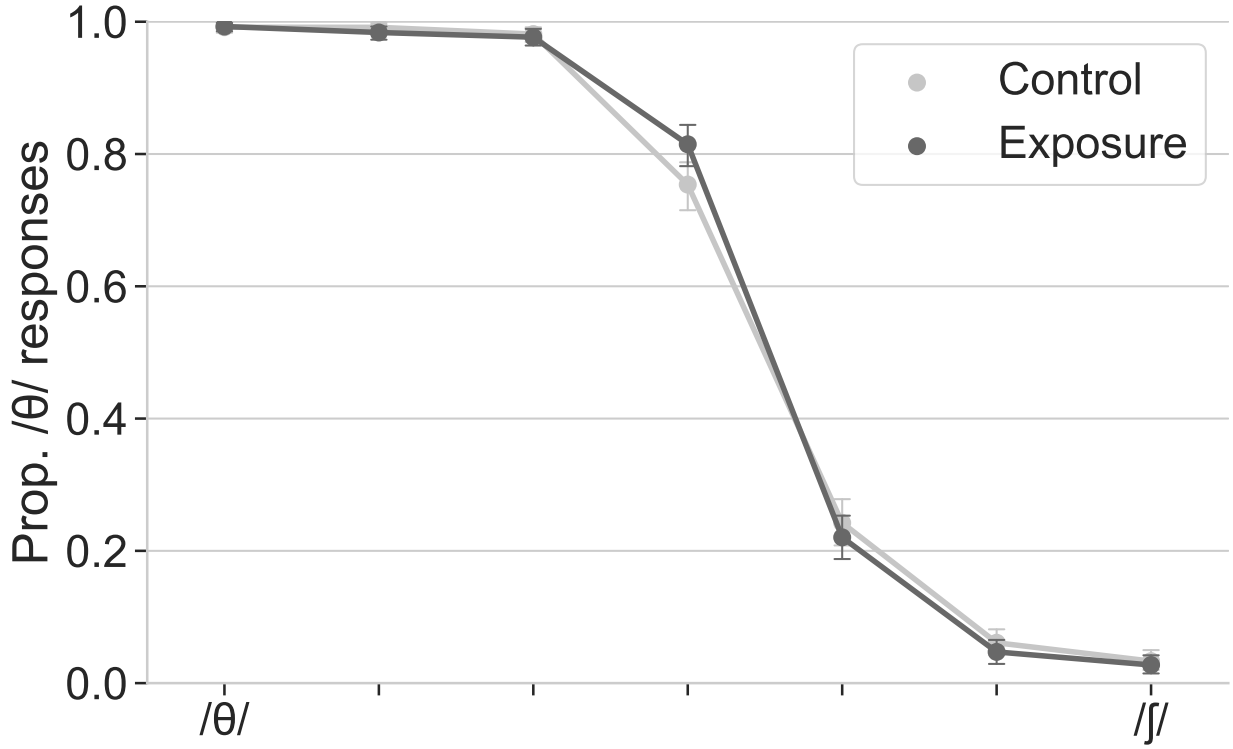


Figure 4.5: Male exposure speaker: proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [ʃ] phonetic continua, by exposure condition.

There were also significant 2-way interactions of step and word position ($\chi^2(1) = 5.73$, $p < 0.01$), of group and block ($\chi^2(1) = 10.43$, $p < 0.001$), and of word position and block ($\chi^2(1) = 15.66$, $p < 0.001$). There were also 3-way interactions of step, group, and word position ($\chi^2(1) = 9.77$, $p < 0.01$), of step, word position, and block ($\chi^2(1) = 6.28$, $p < 0.05$), and a 4-way interaction of step, group, word position, and block ($\chi^2(1) = 5.36$, $p < 0.05$).

The training effect showed up as a significant decrease (-3.2%) in the proportion of /θ/ responses in the data (b = -0.90, SE =0.35, z-value = -2.60 $p < 0.01$) There was also a decrease in the size of the training effect across experiment blocks: the large -5.1% shift seen in the first half of trials decreased to -1.3% in the second half (b = 0.85, SE = 0.26, z-value = 3.23, $p < 0.01$). Similar to the results for [θ]-[s] categorization for this speaker in Experiment 6b, the training effect again seems to hold across continuum steps (see Figure 4.6), rather than being restricted to just the the middle steps.

As expected, the model predicted a lower likelihood of /θ/ responses closer to the /ʃ/ endpoint of the continuum (b = -1.94, SE = 0.17, z-value = -11.29, $p < 0.001$). It also predicted lower overall likelihood of a /θ/ response in the second block of trials (b = -0.86, SE = 0.28, z-value = -4.70, $p < 0.001$), and lower likelihood of a /θ/ response in continua with word-initial /θ/ (b = -2.08, SE = 0.26, z-value = -7.95, $p < 0.001$). The model predicted less

/θ/ responses for the training group, i.e., a shift in categorization responses in the opposite direction typically observed in recalibration studies. Interactions of step and word position suggested a stronger effect of step in word-initial continua (b = -0.33, SE = 0.14, z-value = -2.39, $p < 0.05$), which intensified in the second block of trials (b = -0.53, SE = 0.21, z-value = -2.51, $p < 0.05$). The model also predicted more /θ/ responses for word-initial continua in the second block of trials (b = 1.05, SE = 0.27, z-value = 3.96, $p < 0.001$). The interaction of group and block also suggest that the training effect diminished in the second block of trials (b = 0.85, SE = 0.26, z-value = 3.23, $p < 0.01$). The interaction of step, group, and word position suggest a stronger effect of step for the training group in word-initial continua (b = -0.98, SE = 0.31, z-value = -3.13, $p < 0.01$), but this diminished in the second block of trials (b = 0.94, SE = 0.40, z-value = 2.32, $p < 0.05$).

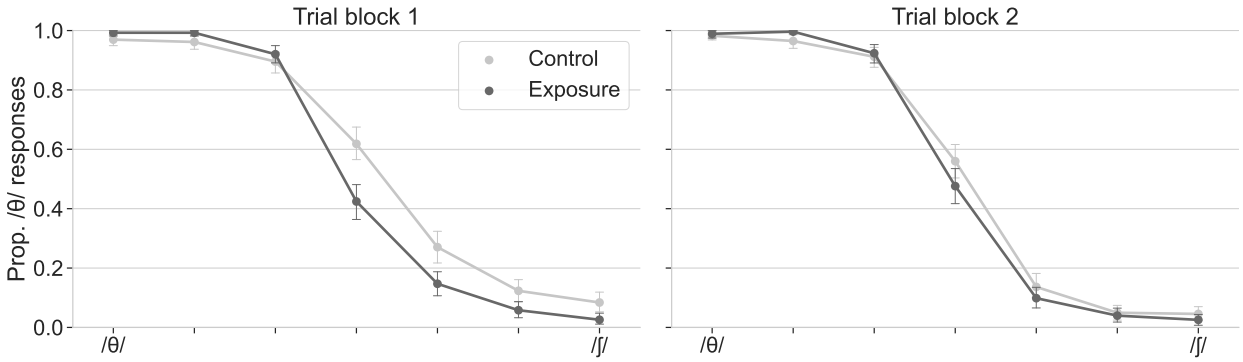


Figure 4.6: Male generalization speaker: proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [ʃ] phonetic continua, by exposure condition and experiment half.

Exp. 7c results

Results showed main effects of continuum step ($\chi^2(1) = 112.44$, $p < 0.001$), experimental group ($\chi^2(1) = 5.01$, $p < 0.05$), and trial block ($\chi^2(1) = 5.21$, $p < 0.05$), with a significant interaction of step and word position ($\chi^2(1) = 5.66$, $p < 0.05$).

The training effect was seen as a significant increase (1.8%) in the proportion of /θ/ responses (b = 1.00, SE = 0.45, z-value = 2.24, $p < 0.05$) There was a numeric decrease in the size of the training effect across experiment blocks (the 2.3% shift seen in the first half of trials decreased to 1.4% in the second half), but this effect did not reach significance ($\chi^2(1) = 2.17$, $p = 0.14$). Similar to the results for the male exposure speaker, the training effect again seems to be localized to just the middle steps of the continuum (see Figure 4.7).

As expected, the model predicted a lower likelihood of /θ/ responses closer to the /ʃ/ end of the continuum (b = -2.72, SE = 0.26, z-value = -10.60, $p < 0.001$), with a stronger effect of step in word-initial continua (b = -0.65, SE = 0.27, z-value = -2.38, $p < 0.05$). A higher likelihood of /θ/ responses was also seen for the second block of trials (b = 0.73, SE = 0.32, z-value = 2.28, $p < 0.05$).

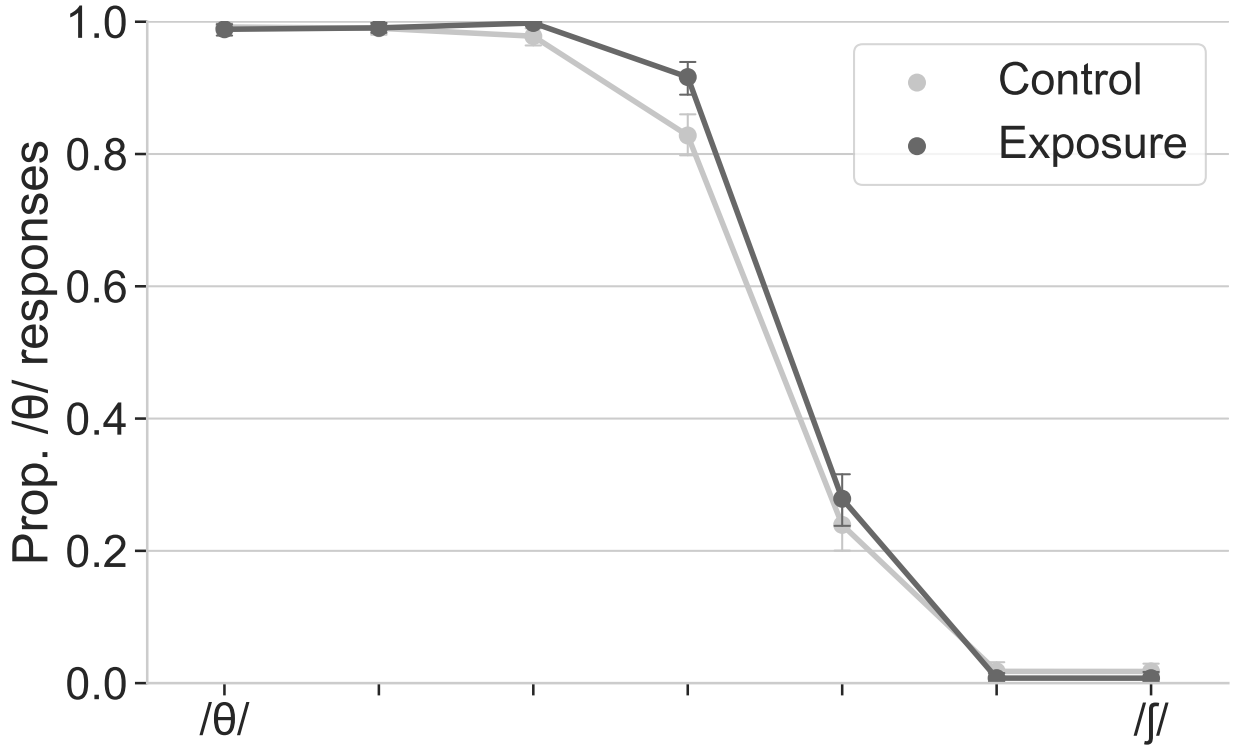


Figure 4.7: Female generalization speaker: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [ʃ] phonetic continua, by exposure condition.

Discussion

Results of this set of experiments are surprising for several reasons. First, there was a failure to replicate the finding in Experiment 2, which showed a shift in categorization of /θ/-/ʃ/ continua following exposure to a /θ/ = [θ/s] accent. This null result may be partially explained by the fact that both the control and training group perceived the test stimuli in a highly categorical way, in contrast with the somewhat more linear response curve seen for the MTurk participants in Experiment 2. In fact, the small numeric category shift in Experiment 7a was restricted entirely to the very middle continuum step, whereas in Experiment 2 it was evident even at the /θ/ endpoint of the continuum. Second, the presence of a significant negative shift (i.e., towards the /θ/ end of the continuum) for the novel male speaker was anomalous and has not been reported in existing recalibration literature. This result suggests that accent exposure led to listeners becoming less tolerant of atypical /θ/ realizations for this speaker when they did not match the training accent. If recalibration caused listeners to expect an /s/-like /θ/, then the perceptual match here may have been poor for this listener group, leading to fewer /θ/ responses. Interestingly, however, participants tested on the novel female speaker did see a significant category shift in the expected direction (towards /ʃ/).

This suggests that perceptual learning transferred to this speaker, suggesting that listeners judged this speaker's productions of /θ/ to be sufficiently similar to that of the exposure speaker to trigger recalibration.

4.6 Acoustic analysis

Given the mixed results obtained in this set of experiments and evidence from previous studies that acoustic similarity is a key predictor of generalization, we conducted acoustic analyses of training and test stimuli. Following Kraljic and Samuel (2005), we obtained center of gravity measures (spectral mean weighted by amplitude) for the middle 75% of target fricatives. This gave a single measurement for each of the 20 critical /θ/ training items, and 56 measurements per speaker for the test items (7 tokens for each of 8 continua across Experiments 6 and 7).

First, results of the training data show an average center of gravity of 4183 Hz for the male exposure speaker. Tokens from the same speaker in the [θ]-[s] test continua (Exp.6a) show a higher mean center of gravity of 5312 Hz. Still, this measure is much lower than for the novel male speaker (6351 Hz) or the novel female speaker (7351 Hz). These results suggest a fairly straightforward explanation of the patterns of generalization obtained in Exp.6 a-c, as they show a continuum of acoustic similarity across the three speakers (see Figure 4.8). First, despite the difference in average spectral means between the exposure speaker's training and test materials, we replicated the significant category boundary shift obtained in Experiment 1. We also saw a large marginally significant shift for the novel male speaker, who was the closest acoustic match to the exposure speaker. Meanwhile, no corresponding shift for the novel female speaker was obtained, consistent with the larger acoustic difference.

For Experiments 7a-c, results are less straightforward. First, spectral means for all speakers' test continua are much closer acoustically. The male exposure speaker's means for [θ]-[ʃ] continua are 4223 Hz, nearly identical to that of the /θ/ = [θ/s] training items (4183 Hz). This close acoustic match may explain why generalization was observed for this contrast in Experiment 2. However, Experiment 7a failed to replicate this result from Experiment 2 (although there was a trend in the expected direction). Meanwhile, the generalization speakers in this set of studies show nearly identical measures: the novel male speaker's test tokens have an average center of gravity of 4702 Hz and the female speaker's tokens are 4681 Hz. This similarity is surprising given opposite categorization results observed for these speakers (the female speaker showed a category shift toward /ʃ/ while the novel male speaker showed a negative shift toward /θ/). Overall, such mixed results suggest that this acoustic measure of fricative similarity may not capture cross-speaker acoustic differences for these sounds very well.

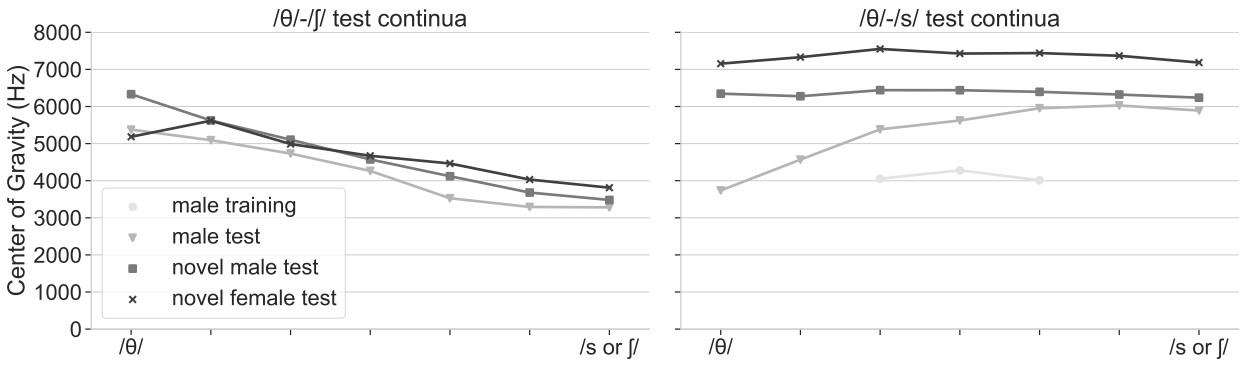


Figure 4.8: Spectral means (center of gravity) for training and test stimuli from Experiments 6-7, by speaker.

4.7 General discussion

Together, the set of results presented in this chapter suggest that generalization of perceptual learning for a novel accent is modulated by several important factors.

The null result in Experiment 5, where no effect of training was observed with any of the speakers (including the familiar male speaker from the exposure phase) indicates that the type of exposure is an important factor determining whether recalibration can be observed in the first place. Interestingly, this finding contrasts with those reported by Reinisch and Holt (2014) and Reinisch et al. (2013), who utilized a similar task where tokens from multiple minimal-pair continua and multiple speakers were intermixed in the test phase. Although they suggest that this kind of methodology can facilitate generalization of learning, we find the opposite effect here. Possibly, the additional variability in the current study (three speakers presented simultaneously rather than two) may have introduced additional uncertainty for listeners, preventing training effects from emerging. Reinisch and Holt (2014) find that the way that listeners' perceived the exposure speaker during the test (categorization) phase changed as a function of the novel (female or male) speaker also presented during exposure. While this did not neutralize the training effect in their study, it is possible that the higher-variability testing paradigm in the current study did. Given that subsequent experiments with single-speaker test phases did show reliable effects of training, this may be an important takeaway for future research on the generalization of phonetic learning.

Results of Experiments 6a-c, consistent with existing literature, show that recalibration of category boundaries for fricatives is largely speaker-specific: listeners trained on a male speaker with an ambiguous pronunciation /θ/ = [θ/s] showed a significant increase in proportion of /θ/ responses during categorization of [θ]-[s] minimal-pair continua. They also showed a marginally significant generalization of phonetic learning to a novel male speaker, but not to a novel female speaker. This pattern of results is consistent with those of Reinisch

and Holt (2014), who initially found transfer of learning only between speakers matched on gender. Generalization from the female training speaker to the novel male speaker was only observed when the authors resampled the male speaker's test stimuli to more closely match the range of phonetic space covered by the female speaker's test stimuli. In the current set of studies, a marginally significant category shift was observed for the more acoustically similar novel male speaker, but not for the more distinct female speaker. This result, again, is consistent with existing work showing that acoustic similarity between speakers is an important predictor of whether learning transfers. This has been previously demonstrated for recalibration studies (Kraljic & Samuel, 2005; Reinisch et al., 2014) as well as for studies utilizing natural accents (Alexander & Nygaard, 2019; Xie, Weatherholtz, et al., 2018).

Results of the second set of experiments (7a-c) are less easily explicable based on the acoustic analyses presented here. Given the acoustically close match between the [θ]-[ʃ] test materials and the training materials (and given the successful within-speaker transfer of learning in Experiment 2), we might have expected to see generalization occur for all speakers. Instead, results show that a category shift for the male exposure speaker trended in the expected direction, and the novel female speaker saw a significant categorization shift, also in the expected direction. However, the novel male speaker showed the opposite shift, seeing a decrease in /θ/ responses across the continuum. This shift cannot be explained by the acoustic analyses presented here, given that this speaker's test materials were nearly identical to those of the female speaker. Moreover, both speakers' test materials were highly similar to the exposure speaker's training materials. In fact, [θ]-[ʃ] test materials were acoustically more similar to the /θ/ = [θ/s] training materials than the [θ]-[s] continua were. This difference is especially puzzling given the results of Experiments 6a-c, which showed the opposite pattern — a numerically large shift for the male generalization speaker, but not for the female speaker. This discrepancy suggests that the spectral measures of acoustic similarity utilized here, although previously used to explain differences in cross-speaker generalization of phonetic learning (Kraljic & Samuel, 2005) may be insufficient to capture differences between speakers in the present study. Exploring additional acoustic measures (e.g., RMS amplitude) that have been shown to distinguish sibilant and non-sibilant fricatives (Jongman et al., 2000) could shed light on this question.

Overall, these results indicate, consistent with previous work, that recalibration of phonetic categories can affect perception of a novel speakers (Kraljic & Samuel, 2005, 2006; Reinisch & Holt, 2014). However, it is unclear from the experiments presented in this chapter what the mechanisms for generalization of such learning are, given the tenuous relationship with the pattern of results observed in Experiments 1-2. For the male exposure speaker, a significant shift for [θ]-[s] continua but not [θ]-[ʃ] continua shows a failure to replicate the pattern of results found in Chapter 1. While unexpected, these results suggest a relatively targeted learning mechanism that is specific for the trained contrast and does not result in changes to perception of neighboring sounds. For the novel male speaker, results also support a relatively specific learning strategy. Trained listeners showed a marginally significant shift toward /s/ when tested on [θ]-[s] continua, and a shift toward /θ/ when tested on /θ/-/ʃ/ continua. This suggests an increased tolerance for atypical pronunciations

when these matched the exposure accent, coupled with a decrease in tolerance for atypical pronunciations of /θ/ that did not match the exposure accent. The pattern of results observed across experiments for the novel female speaker is less clear. The lack of a shift on [θ]-[s] continua but a significant shift toward /ʃ/ for /θ/-/ʃ/ continua could be explained by acoustic similarity: analysis shows that this speaker's test items for the [θ]-[ʃ] continua were acoustically much closer to the exposure speaker's training items than her /θ/-/ʃ/ continua. However, an acoustically-based explanation here would also predict a similar shift on /θ/-/ʃ/ continua for both the familiar and novel male speakers. Overall, however, the pattern of results across experiments in this chapter supports a relatively specific learning mechanism. No speaker saw a significant increase in /θ/ responses for both /θ/-/s/ and /θ/-/ʃ/ continua, which we would expect under a category expansion mechanism. In fact, results for the novel male speaker speaker suggest a much more conservative learning mechanism shows that listeners were *less* likely to categorize /ʃ/-like tokens as examples of /θ/, suggesting more stringent categorization criteria for instances of this sound that did not match the training accent. This suggests that different mechanisms may underlie the generalization of phonetic recalibration effects than those which we see within-speaker.

The mixed results in this chapter may not be entirely surprising given the conflicting pattern of results in the broader literature on perceptual learning for accented speech. In particular, generalization following single-speaker accent exposure is not inconsistent across studies. Bradlow and Bent (2008) found that single-speaker exposure to Mandarin-accented English yielded no benefit in comprehension of a novel speaker with the same accent. Weil (2001) did find generalization to a novel speaker, but transfer of learning was not consistent across tasks (e.g., it was observed with sentence level but not word-level tasks). Witteman et al. (2013) found that perceptual learning of German-accented Dutch could transfer to a novel speaker, but not immediately after exposure (benefits to lexical processing only emerged in the second half of the experiment). Finally, Xie and Myers (2017) found that, regardless of whether listeners received single-speaker or multi-speaker accent familiarization, transfer of phonetic learning for Mandarin-accented English was predicted by acoustic similarity between familiarized and novel speakers. This complements findings of Alexander and Nygaard (2019), who found inconsistent generalization with different types of multi-accent training — instead, the key predictor of generalization in their study was between-speaker acoustic similarity.

The common thread through much of the research on generalization of phonetic learning for accented speech is the importance of acoustic similarity in predicting generalization. The set of experiments here provide some support for the importance of this factor as a predictor of generalization in Experiments 6a-c, but fails to account for the mixed findings seen in Experiments 7a-c. Given the relative coarseness of fricative spectral means as a sole measure of acoustic similarity, it is plausible that this listeners were relying on other measures of acoustic and/or perceptual similarity in determining whether and how learning should generalize to novel speakers. Future studies may shed more light on which acoustic factors listeners attend to in determining whether two speakers are sufficiently similar for learning to generalize.

Chapter 5

Conclusion

5.1 Mechanisms of accent adaptation

The primary goal of this dissertation was to investigate the mechanism(s) underlying perceptual learning of accented speech. Specifically, the series of experiments in this study investigate three largely open questions in this literature. The first asks what mechanisms underlie the recalibration or 'retuning' of phonetic category boundaries following exposure to an atypical pronunciation. The second asks whether these same mechanisms affect subsequent online processing of novel words produced by the same speaker. The third question asks whether the same mechanisms that underlie phonetic learning within a single speaker also constrain the transfer of learning to novel speakers.

Each of these points are addressed in turn in this dissertation. Chapter 2 addresses an ambiguity in the perceptual recalibration literature. These experiments test whether recalibration of perceptual category boundaries following exposure to an atypical pronunciation is caused by a targeted mechanism specific to the exposure accent (category shift) or a more general mechanism of criteria relaxation (category expansion). Given that studies of natural accent accommodation have found evidence of criteria relaxation following accent exposure (e.g., Zheng and Samuel, 2020), we might expect this mechanism to also drive changes in internal category structure. The experiments in this chapter provide evidence for limited (non-uniform) category expansion that appears to be constrained by perceptual similarity, supporting a possible link between recalibration and accent accommodation. Chapter 3 investigates whether the same kind of mechanism could benefit online lexical processing following exposure to a speaker with an atypical pronunciation. Results of this study are consistent with a category expansion mechanism, although the degree to which listeners tolerate novel accented productions of a given target sound appears to be constrained by the accent exposure context and perceptual similarity to the exposure accent. Finally, Chapter 4 investigates whether the same mechanism driving recalibration of category boundaries for a single speaker also applies to novel speakers. Results of this set of experiments provide evidence for some generalization of phonetic recalibration to both novel speakers and novel

phonetic contrasts (as in Chapter 2), but the pattern of results does not clearly align with those obtained in Chapter 2 and appears to vary across speakers. On the whole, however, these results favor a more conservative mechanism for generalization of phonetic category retuning.

Taken together, these results show that recalibration of phonetic category boundaries meets a number of criteria supporting a relationship between changes to category structure and accent accommodation. First, a relatively general mechanism of phonetic retuning, as supported by results from Chapters 2 and 3, may be beneficial to accent accommodation. Given the increased variability of accented speech (Wade et al., 2007), it may behoove listeners to main a larger perceptual "window" for sound categorization than they otherwise would. For instance, studies have shown that non-native speakers of English may produce /θ/ in different ways (e.g., substituting /f/, /s/, or /t/) even when they speak the same L1. Thus, the optimal strategy for perceptually adapting to a given accent may be to relax categorization criteria to accommodate the multiple possible variants of a given category that a listener might encounter for a given accent. Crucially, however, an unconstrained expansion of category boundaries may not be the best perceptual strategy, given the systematic regularities typically seen within a given accent and evidence that listeners adapt to these regularities (Hanulíková & Weber, 2012). A second point in favor of a relationship between category structure and accent accommodation is the finding that this same mechanism (category expansion) appears to constrain subsequent lexical processing of the same speaker, as has been previously found with naturally-accented speech (Xie et al., 2017). Finally, the results on generalization of phonetic recalibration to novel speakers in this dissertation indicate that listeners may use multiple criteria to determine whether learning generalizes. Given the differences in patterns of generalization across speakers, it seems that listeners may rely on more targeted mechanisms when generalizing phonetic learning.

Crucially, while the experiments presented here tentatively favor a category expansion explanation for both recalibration of category boundaries (Chapter 2) and changes in lexical processing (Chapter 3), they do not exclude a more targeted mechanism as a possibility for accent adaptation. As noted in this literature review, there is evidence for both types of mechanisms in the existing literature on perceptual learning for accented speech. Previous research has also highlighted the possibility that both mechanisms could be on the table. For instance, Schmale et al. (2012) suggest that a category expansion strategy may be used in the early stages of accent learning, but that with sufficient exposure listeners adopt a more targeted strategy that is specific to regularities in the pronunciation. Given the short timescales on which accent adaptation occurs (sometimes within minutes), it is plausible that a form of category expansion may be the preferred initial strategy. It remains unclear whether the kind of adaption that occurs with long-term exposure to accented speech may involve a fundamentally different set of perceptual adaptations.

5.2 Important factors in phonetic learning for accented speech

An important goal of this dissertation has been to bridge two related bodies of literature on perceptual learning: recalibration of phonetic category boundaries vs. natural accent accommodation. These different approaches to studying perceptual learning for speech, despite the common assumption that they are linked, serve as extreme endpoints on the spectrum of methodologies for studying atypical speech. On the one hand, recalibration work probably represents the least veridical approximation of the real-world task of accent accommodation: these studies rely on a single perfectly ambiguous target sound in a single otherwise natively-accented voice. Perhaps for this reason, this approach to studying perceptual learning for speech has been a popular one, given the ease of experimentally specifying the parameters of deviation from a "normal" pronunciation. The recalibration paradigm also offers a straightforward way to study the mechanisms that listeners use to adapt to accented speech, since perceptual learning can be evaluated at the phonemic (or allophonic) level with a relatively high degree of specificity. On the other hand, studies of natural accent accommodation commonly use a more naturalistic set of tasks during exposure and test, which more closely approximate real-world learning contexts. This approach also has its problems, however, since the types of measures often used to assess learning (e.g, comprehension scores or fluency of lexical processing) are often too coarse to give much insight into the mechanisms involved. At the same time, because non-native accents typically deviate in multiple ways from natively-accented speech (e.g., showing differences in duration, prosody, segmental realization, etc.), it is often difficult to evaluate exactly *what* it is that listeners are learning during accent adaption.

This dissertation adopts the recalibration paradigm for the reasons outlined above, while recognizing the inherent limitations of this approach. Critically, we assume that accent adaptation occurs at least in part at the segmental level, and that changes in segmental representations allow for word-independent learning of the systematic phonetic patterns that characterize non-natively accented speech. This results in reduced mismatch as listeners adjust their expectations for how a given sound may be pronounced (Clark, 2013). While there is evidence for such segment-level adaptation to naturally-accented speech (e.g., Witteman et al., 2013; Xie and Myers, 2017; Xie et al., 2017), it is likely that listeners also adapt in other ways. For instance, Reinisch and Weber (2012) found that listeners were able to rapidly adapt to lexical stress errors in non-natively accented speech. Lev-Ari (2015) suggests that processing of non-natively accented speech involves a fundamentally different listening mode — because listeners expect non-native speakers to be less reliable/predictable, they instead focus more on using top-down context (i.e., non-linguistic context) to process speech. That is, they place less weight on the speech signal itself and more on the speech context (e.g., their prior expectations about what a non-native speaker intends to convey, rather than what they actually say). An ERP study by Hanulíková et al. (2012), for instance, found that grammatical gender violations in Dutch elicited a P600 response (indicating syntactic

anomaly) when occurring in natively-accented speech, but not in a non-native accent. This suggests that listeners may expect and tolerate more errors in non-native speech, and that this occurs at multiple levels of linguistic processing.

Crucially, the nature of perceptual adaptation may depend on multiple factors. For instance, important factors for generalization of perceptual learning for naturally-accented speech have been shown to include sufficient input variability (e.g., multi-speaker and/or multi-accent exposure) as well as between-speaker acoustic similarity. An important limitation of this study (and other typical recalibration studies) is the use of a single-speaker exposure phase — this may not always provide sufficient input variability for robust generalization. Another important factor is the nature of exposure and test stimuli. Previous work has shown a discrepancy between perceptual learning at the word vs. sentence level Weil (2001). This suggests that accent learning may involve adaptations beyond the segmental level, so single-word exposure may be insufficient for listeners to learn relevant differences in prosody or stress that characterize given accent. Given the focus of the current study on the mechanisms of segment-level learning, it necessarily misses these potentially important aspects of accent adaptation. This means we should be cautious when drawing comparisons to other studies that have used measures such as comprehension scores, since improved comprehension could be driven by a stronger reliance on top-down context rather than phonetic learning (Lev-Ari, 2015).

5.3 Stability vs. plasticity in perceptual learning

As we have seen, an important question in the general literature on perceptual learning is how perceivers solve the stability-plasticity dilemma; simultaneously being able to adapt to novel information in the environment without forgetting what they have already learned (Fahle, 2005; Grossberg, 2005; Kleinschmidt & Jaeger, 2015). In the case of perceptual learning for speech, one answer to this question appears to be that listeners 'retune', or 'recalibrate' linguistic representations to better match recently encountered input. Importantly, the current study shows that listeners can rapidly adapt to a novel pronunciation, showing evidence of changes in both internal phonetic category structure and subsequent lexical processing. Moreover, such learning appears to generalize beyond the specific training stimuli to novel phonetic contrasts, novel words, and novel speakers. This finding aligns with the observation in the general literature on perceptual learning, which shows that perceptual learning for complex tasks occurs rapidly and generalizes easily. Given the highly variable nature of the speech signal, such rapid and flexible learning appears to be an important adaptation of the perceptual system that allows for relatively effortless communication across diverse speech contexts.

Bibliography

Alexander, J. E. D., & Nygaard, L. C. (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *145*(6), 3382–3398. https://doi.org/10.1121/1.5110302

Apps, M. A. J., & Tsakiris, M. (2013). Predictive codes of familiarity and context during the perceptual learning of facial identities. *Nature Communications*, *4*(1), 2698. https://doi.org/10.1038/ncomms3698

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180. https://doi.org/10.1121/1.4789864

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00328

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv:1406.5823 [stat]*.

Bent, T., & Holt, R. (2013). The influence of talker and foreign-accent variability on spoken word identification. *The Journal of the Acoustical Society of America*, *133*(3), 1677–1686. https://doi.org/10.1121/1.4776212

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies [Publisher: Springer]. *Behavior research methods*, *44*(4), 991–997.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Charoy, J. (2021). *Accommodation to non-native accented speech: Is perceptual recalibration involved?* (Doctoral dissertation). State University of New York at Stony Brook.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Cooke, M., & García Lecumberri, M. L. (2021). How reliable are online speech intelligibility studies with known listener cohorts? *The Journal of the Acoustical Society of America*, *150*(2), 1390–1401. https://doi.org/10.1121/10.0005880

Cooper, A., & Bradlow, A. (2018). Training-induced pattern-specific phonetic adjustments by first and second language listeners. *Journal of Phonetics*, *68*, 32–49. https://doi.org/10.1016/j.wocn.2018.02.002

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00479

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of english phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678. https://doi.org/10.1121/1.1810292

Diehl, R. L. (1989). Feature detectors for speech: A critical reappraisal. *Psychonomic Bulletin*, *89*(1), 1–18.

Diehl, R. L., Elman, J. L., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 599–609.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*(1), 99–109. https://doi.org/10.1016/0010-0285(73)90006-6

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time (l). *J. Acoust. Soc. Am.*, *119*(4), 4.

Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00148

Fahle, M. (2001). Perceptual learning. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (1st ed). Elsevier.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology*, *15*(2), 154–160. https://doi.org/10.1016/j.conb.2005.03.010

Fahle, M. (2009). Perceptual learning and sensory plasticity. In *Encyclopedia of neuroscience* (pp. 523–533). Elsevier. https://doi.org/10.1016/B978-008045046-9.00230-8

Fine, I., & Jacobs, R. A. (2002). Comparing perceptual learning across tasks: A review. *Journal of Vision*, *2*(2), 5–5. https://doi.org/10.1167/2.2.5

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency [Publisher: Nature Publishing Group]. *Nature*, *287*(5777), 43–44.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–3134. https://doi.org/10.1121/1.413041

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

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? [Publisher: American Psychological Association]. *Psychological review*, *62*(1), 32.

Gilbert, C. D., Li, W., & Piech, V. (2009). Perceptual learning and adult cortical plasticity: Perceptual learning and adult cortical plasticity. *The Journal of Physiology*, *587*(12), 2743–2751. https://doi.org/10.1113/jphysiol.2009.171488

Gold, J. I., & Watanabe, T. (2010). Perceptual learning. *Current Biology*, *20*(2), R46–R48. https://doi.org/10.1016/j.cub.2009.10.066

Goldstone, R. L. (1998). Perceptual learning [Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA]. *Annual review of psychology*, *49*(1), 585–612.

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Schurman, J. (2010). Short-term adaptation to accented english by younger and older adults. *The Journal of the Acoustical Society of America*, *128*(4), EL200–EL204. https://doi.org/10.1121/1.3486199

Grossberg, S. (2005, January 1). Linking attention to learning, expectation, competition, and consciousness. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 652–662). Academic Press. https://doi.org/10.1016/B978-012375731-9/50111-7

Hall, G. (2008). Perceptual learning. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (pp. 103–121). Elsevier.

Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person s mistake is another s standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–87.

Hanulíková, A., & Weber, A. (2012). Sink positive: Linguistic experience with th substitutions influences nonnative word recognition. *Attention, Perception, & Psychophysics*, *74*(3), 613–629. https://doi.org/10.3758/s13414-011-0259-7

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, *62*(4), 874–886. https://doi.org/10.3758/BF03206929

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*(5), 943–950. https://doi.org/10.3758/s13423-011-0129-2

Johnson, K. (2003). *Acoustic and auditory phonetics*. Blackwell Publishing.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252. https://doi.org/10.1121/1.1288413

Kawahara, H., & Morise, M. (2011). Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana*, *36*(5), 713–727. https://doi.org/10.1007/s12046-011-0043-3

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. https://doi.org/10.1037/a0038695

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81. https://doi.org/10.1016/j.cognition.2007.07.013

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268. https://doi.org/10.3758/BF03193841

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Leikin, M., Ibrahim, R., Eviatar, Z., & Sapir, S. (2009). Listening with an accent: Speech perception in a second language by late bilinguals. *Journal of Psycholinguistic Research*, *38*(5), 447–457. https://doi.org/10.1007/s10936-009-9099-1

Leonards, U., Rettenbach, R., Nase, G., & Sireteanu, R. (2002). Perceptual learning of highly demanding visual search tasks. *Vision Research*, *42*(18), 2193–2204. https://doi.org/10.1016/S0042-6989(02)00134-7

Lev-Ari, S. (2015). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01546

Li, W., & Gilbert, C. (2009, January 1). Perceptual learning: Neural mechanisms. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 535–541). Academic Press. https://doi.org/10.1016/B978-008045046-9.00227-8

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1562–1588. https://doi.org/10.1037/xhp0000693

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [sh]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228.

Marslen-Wilson, W., Moss, H. E., College, B., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1376–92.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. https://doi.org/10.1080/03640210802035357

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126. https://doi.org/10.1207/s15516709cog0000_79

Melguy, Y. V., & Johnson, K. (2021). General adaptation to accented english: Speech intelligibility unaffected by perceived source of non-native accent. *The Journal of the Acoustical Society of America*, *149*(4), 2602–2614. https://doi.org/10.1121/10.0004240

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352. https://doi.org/10.1121/1.1907526

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512. https://doi.org/10.3758/BF03208147

Mitterer, H., Cho, T., & 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. https://doi.org/10.1016/j.wocn.2016.03.001

Mitterer, H., & Reinisch, E. (2017). Surface forms trump underlying representations in functional generalisations in speech perception: The case of german devoiced stops. *Language, Cognition and Neuroscience*, *32*(9), 1133–1147. https://doi.org/10.1080/23273798.2017.1286361

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361. https://doi.org/10.1016/j.cognition.2013.07.011

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800305

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 3.3.2). Vienna, Austria, R Foundation for Statistical Computing.

Reinisch, E., & 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), 539–555. https://doi.org/10.1037/a0034409

Reinisch, E., & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics*, *55*, 96–108. https://doi.org/10.1016/j.wocn.2015.12.004

Reinisch, E., & Weber, A. (2012). Adapting to suprasegmental lexical stress errors in foreign-accented speech. *The Journal of the Acoustical Society of America*, *132*(2), 1165–1176. https://doi.org/10.1121/1.4730884

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 75–86. https://doi.org/10.1037/a0027979

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics*, *45*, 91–105. https://doi.org/10.1016/j.wocn.2014.04.002

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure: Brief exposure to an unfamiliar accent. *Developmental Science*, *15*(6), 732–738. https://doi.org/10.1111/j.1467-7687.2012.01175.x

Schuhmann, K. S. (2016). Cross-linguistic perceptual learning in advanced second language listeners. *Proceedings of the Linguistic Society of America*, *1*, 31. https://doi.org/10.3765/plsa.v1i0.3731

Schuhmann, K. S. (2014). Perceptual learning in second language learners [Publisher: State University of New York at Stony Brook].

Seibert, A. (2011). *A sociophonetic analysis of l2 substitution sounds of american english interdental fricatives* (Master's thesis). Southern Illinois University at Carbondale.

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 67–113). McGraw-Hill.

Sidaras, S. K., Alexander, J. E. D., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306. https://doi.org/10.1121/1.3101452

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

Strand, E. A., & Johnson, K. (1996, December 31). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology* (pp. 14–26). De Gruyter. https://doi.org/10.1515/9783110821895-003

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–136. https://doi.org/10.1016/j.cognition.2010.10.018

Vaughn, C. R. (2019). Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America*, *145*(5), 3218–3232. https://doi.org/10.1121/1.5108831

Volkmann, A. W. (1858). Uber den einfluss der uebung. *Leipzig Berichte Math.-phys. Classe*, *10*, 38–39.

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2), 122–144. https://doi.org/10.1159/000107913

Weatherholtz, K. (2015). *Perceptual learning of systemic cross-category vowel variation* (Doctoral dissertation). The Ohio State University.

Weil, S. (2001). Foreign accented speech: Encoding and generalization. *The Journal of the Acoustical Society of America*, *109*(5), 2473–2473. https://doi.org/10.1121/1.4744779

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, *35*(1), 49–64. https://doi.org/10.3758/BF03205924

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers: Toddler accent adaptation. *Developmental Science*, *14*(2), 372–384. https://doi.org/10.1111/j.1467-7687.2010.00986.x

Witteman, M. J., Bardhan, N. P., Weber, A., & McQueen, J. M. (2015). Automaticity and stability of adaptation to a foreign-accented speaker. *Language and Speech*, *58*(2), 168–189. https://doi.org/10.1177/0023830914528102

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, *75*(3), 537–556. https://doi.org/10.3758/s13414-012-0404-y

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210. https://doi.org/10.1080/23273798.2017.1369551

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, *97*, 30–46. https://doi.org/10.1016/j.jml.2017.07.005

Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 206–217. https://doi.org/10.1037/xhp0000285

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, *143*(4), 2013–2031. https://doi.org/10.1121/1.5027410

Zheng, Y., & Samuel, A. G. (2020). The relationship between phonemic category boundary changes and perceptual adjustments to natural accents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(7), 1270–1292. https://doi.org/10.1037/xlm0000788

Appendix A

Chapter 2 materials

A.1 Experiment 1

Table A.1: Critical and filler words used in the lexical decision exposure task (Exp.1 and 2).

Words					Non-words				
/θ/ words	*/s/ words*	*Fillers*							
anthem	reconcile	multitude	aluminum	kimono	adgendoy	dioryle	oudrenoa	hartacko	ayarbik
apathetic	eraser	polymer	undertaker	defending	akelen	udanaco	pelade	pelnimated	nererant
apathy	rigorous	topical	maternal	outnumber	altartalized	dynrem	pleope	bulerame	bonimaded
beneath	clandestine	durable	domination	Carribean	altercole	elember	potler	toalbinade	contaluow
breakthrough	admissible	undertow	mutilated	parachute	tulable	tonker	pocorome	nomikord	odanatar
commonwealth	disparaged	untoward	detachment	commoner	amahate	etoced	polacual	caltacate	tumbodel
Dorothy	hallucinate	workable	gunpowder	broadway	ampoter	etugant	premetor	okenel	cumpamer
earthquake	intelligence	challenged	abandoned	engineer	anapt	guncore	prodabanga	pontradashing	odecogo
empathy	episode	pretended	marketable	optional	ancarrunt	haderate	radorcattoon	cayarac	nobelake
hypothetical	elements	unarmed	compilation	termination	ancorrack	pabutda	rapombargad	coerpage	dadratar
Jonathan	democracy	adorable	degree	moderate	darkackood	andegaul	reatonape	arunimung	oggander
marathon	pregnancy	determine	opportune	dominated	anguilder	horabtalane	reibonairo	cleniot	dargora
Neanderthal	narcotics	underwater	independent	laundry	nantor	kedoac	relecker	puroucly	omblatal
tablecloth	Tennessee	abortion	argument	ambush	noda	lampunger	garempit	umberpater	rallad
telepathy	absence	awarded	unlucky	coronation	becoor	leggarer	klepentip	collattar	omblegon
Timothy	outsmart	terminated	uploading	Kennedy	bocowlable	meloded	ungarnet	nannotad	demurea
twentieth	Arkansas	coordinate	retrograde	commendation	conkartat	mebable	tepelnim	nadelmar	omparkandar
unauthorized	amnesty	opener	bureaucrat	hibernate	malatad	molorat	umbelyaper	combeter	dentakter
undergrowth	prosper	monitored	abnormal	mutilation	booktugner	morachable	abolaper	negryhad	otler
unethical	participate	congregation	Abraham	recommended	rorana	motolad	adorshem	ponimashum	connontnor

Table A.2: Minimal pairs used to generate test continua in Experiment 1.

Group	Minimal pair	Target word-position
A	oath - oaf	word-final
	death - deaf	word-final
	thin - fin	word-initial
	thought - fought	word-initial
B	mouth - mouse	word-final
	math - mass	word-final
	thigh - sigh	word-initial
	think - sink	word-initial

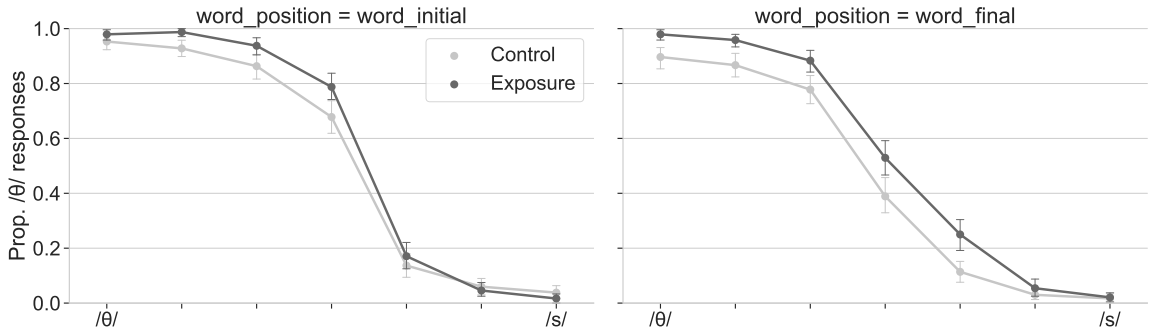


Figure A.1: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [s] phonetic continua, by exposure condition and target word position (Experiment 1).

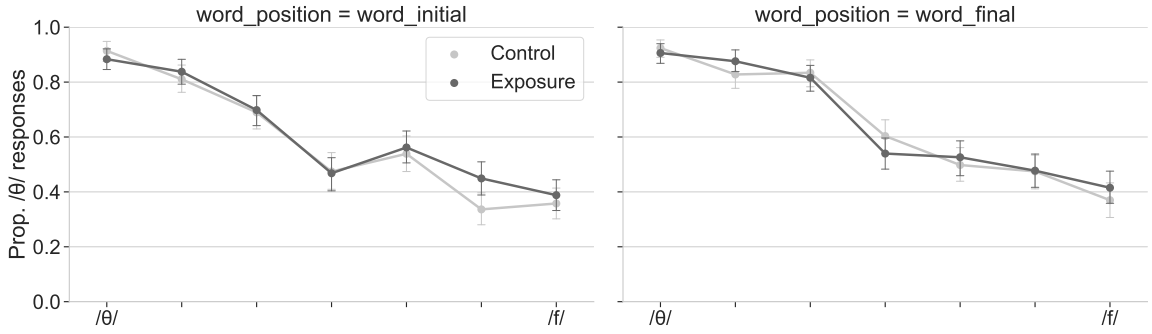


Figure A.2: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [f] phonetic continua, by exposure condition and target word position (Experiment 1).

Experiment 1a (/θ/-/s/ categorization)

Table A.3: Model estimates for Experiment 1a logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * /θ/ word position + group + (1 + center(step) * /θ/ word position | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	-0.239	0.24	-0.99	0.322
center(step)	-2.085	0.14	-15.37	< 2e-16
word_posword_initial	1.075	0.19	5.63	1.8e-08
grouptraining_th-s	0.544	0.27	2.03	0.042
center(step):word_posword_initial	-0.775	0.2	-3.96	7.4e-05

Table A.4: Analysis of Deviance Table (Type II Wald chisquare tests) for Experiment 1a logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	0.981	1	0.322
center(step)	236.327	1	< 2e-16
word_pos	31.711	1	1.8e-08
group	4.139	1	0.042
center(step):word_pos	15.708	1	7.4e-05

Experiment 1b (/θ/-/f/ categorization)

Table A.5: Model estimates for Experiment 1b logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * /θ/ word position + group + (1 + center(step) * /θ/ word position | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	0.821	0.13	6.38	1.8e-10
center(step)	-0.544	0.04	-14.37	< 2e-16
word_posword_initial	-0.326	0.12	-2.72	0.0064
grouptraining_th-f	0.013	0.16	0.09	0.9316
center(step):word_posword_initial	0.074	0.04	1.69	0.0910

Table A.6: Analysis of Deviance Table (Type II Wald chisquare tests) for Experiment 1a logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	40.646	1	1.8e-10
center(step)	206.359	1	< 2e-16
word_pos	7.422	1	0.0064
group	0.007	1	0.9316
center(step):word_pos	2.856	1	0.0910

A.2 Experiment 2

Table A.7: Minimal pairs used to generate test continua in Experiment 2.

Minimal pair	Target word position
math - mash	word-final
wrath rash	word-final
thought - shot	word-initial
thin - shin	word-initial

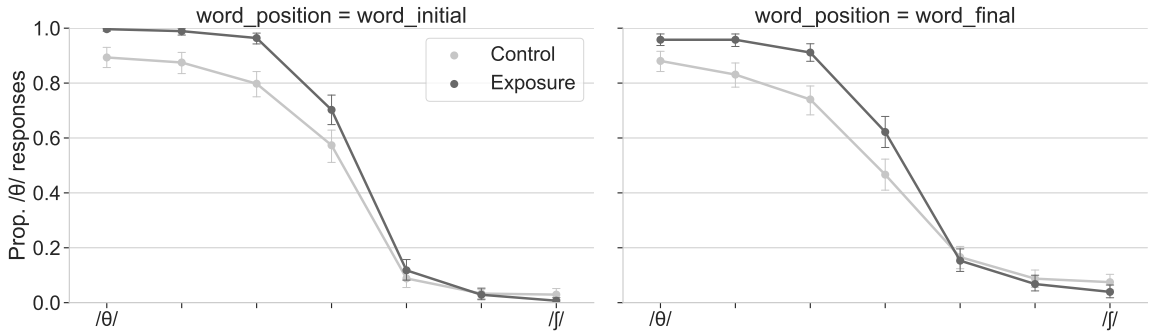


Figure A.3: Proportion /θ/ responses for groups tested on categorizing 7-step [θ] – [ʃ] phonetic continua, by exposure condition and target word position (Experiment 2).

Experiment 2 (/θ/-/ʃ/ categorization)

Table A.8: Model estimates for Experiment 2 logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * /θ/ word position + center(step) * group + (1 + center(step) * /θ/ word position | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	-0.202	0.19	-1.05	0.2943
center(step)	-1.637	0.18	-9.06	< 2e-16
word_posword_initial	0.506	0.17	2.96	0.0030
grouptraining	0.79	0.25	3.16	0.0016
center(step):word_posword_initial	-0.935	0.19	-4.85	1.2e-06
center(step):grouptraining	-0.703	0.26	-2.72	0.0066

Table A.9: Analysis of Deviance Table (Type II Wald chisquare tests) for Experiment 2 logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	1.1	1	0.2943
center(step)	82.085	1	< 2e-16
word_pos	8.78	1	0.0030
group	9.966	1	0.0016
center(step):word_pos	23.513	1	1.2e-06
center(step):group	7.379	1	0.0066

Appendix B

Chapter 3 materials

B.1 Experiment 3

Table B.1: Model estimates for Experiment 3 reaction time analysis. Linear mixed model fit by REML. t-tests use Satterthwaite's method. Formula: scale(log(rt)) ~ condition * group * block + (1 | target) + (1 + block + condition | subj).

	Coeff.	Std. Error	t-value	p-value
(Intercept)	0.357	0.11	3.19	0.00201
conditionidentity	-0.401	0.08	-5.12	1e-06
conditionrelated	-0.238	0.07	-3.51	0.00056
groupexposure	-0.004	0.15	-0.02	0.98118
block2	-0.183	0.08	-2.4	0.01729
conditionidentity:groupexposure	-0.127	0.11	-1.14	0.25444
conditionrelated:groupexposure	0.136	0.1	1.42	0.15860
conditionidentity:block2	-0.077	0.09	-0.84	0.39840
conditionrelated:block2	0.003	0.09	0.04	0.97198
groupexposure:block2	-0.081	0.11	-0.75	0.45544
conditionidentity:groupexposure:block2	0.171	0.13	1.33	0.18342
conditionrelated:groupexposure:block2	-0.034	0.13	-0.26	0.79410

Table B.2: Model estimates for Experiment 3 reaction time analysis (condition variable releveled, ref = "identity"). Linear mixed model fit by REML. t-tests use Satterthwaite's method. Formula: scale(log(rt)) ~ condition * group * block + (1 | target) + (1 + block + condition | subj).

	Coeff.	Std. Error	t-value	p-value
(Intercept)	-0.045	0.12	-0.37	0.70911
conditionunrelated	0.401	0.08	5.12	1e-06
conditionrelated	0.163	0.08	2.13	0.03539
groupexposure	-0.131	0.16	-0.81	0.42072
block2	-0.259	0.08	-3.44	0.00071
conditionunrelated:groupexposure	0.127	0.11	1.14	0.25444
conditionrelated:groupexposure	0.263	0.11	2.42	0.01669
conditionunrelated:block2	0.077	0.09	0.84	0.39840
conditionrelated:block2	0.08	0.09	0.89	0.37472
groupexposure:block2	0.091	0.11	0.85	0.39796
conditionunrelated:groupexposure:block2	-0.171	0.13	-1.33	0.18342
conditionrelated:groupexposure:block2	-0.205	0.13	-1.61	0.10824

Table B.3: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 3 reaction time model

	Chi-squared	Deg. Freedom	p-value
(Intercept)	10.19	1	0.0014
condition	28.014	2	8.3e-07
group	0.001	1	0.9811
block	5.752	1	0.0165
condition:group	6.085	2	0.0477
condition:block	1.005	2	0.6051
group:block	0.559	1	0.4547
condition:group:block	2.957	2	0.2280

B.2 Experiment 4

Table B.4: Model estimates for Experiment 4 reaction time analysis. Linear mixed model fit by REML. t-tests use Satterthwaite's method. Formula: scale(log(rt)) ~ condition * group * block + (1 | target) + (1 + block + condition | subj).

	Coeff.	Std. Error	t-value	p-value
(Intercept)	0.211	0.12	1.8	0.07597
conditionidentity	-0.254	0.06	-4.19	4e-05
conditionrelated	-0.033	0.06	-0.53	0.59916
groupexposure	0.217	0.16	1.34	0.18525
block2	-0.141	0.07	-2.07	0.03944
conditionidentity:groupexposure	-0.071	0.09	-0.82	0.41270
conditionrelated:groupexposure	-0.178	0.09	-1.97	0.04987
conditionidentity:block2	-0.284	0.08	-3.49	0.00049
conditionrelated:block2	-0.225	0.08	-2.78	0.00543
groupexposure:block2	-0.046	0.1	-0.47	0.63725
conditionidentity:groupexposure:block2	0.193	0.12	1.64	0.10082
conditionrelated:groupexposure:block2	0.131	0.12	1.12	0.26402

Table B.5: Model estimates for Experiment 4 reaction time analysis (condition variable releveled, ref = "identity"). Linear mixed model fit by REML. t-tests use Satterthwaite's method. Formula: scale(log(rt)) ~ condition * group * block + (1 | target) + (1 + block + condition | subj).

	Coeff.	Std. Error	t-value	p-value
(Intercept)	-0.043	0.12	-0.35	0.72711
conditionunrelated	0.254	0.06	4.19	4.0e-05
conditionrelated	0.221	0.06	3.79	0.00019
groupexposure	0.146	0.17	0.86	0.39207
block2	-0.425	0.07	-6.24	2.1e-09
conditionunrelated:groupexposure	0.071	0.09	0.82	0.41270
conditionrelated:groupexposure	-0.107	0.08	-1.28	0.20016
conditionunrelated:block2	0.284	0.08	3.49	0.00049
conditionrelated:block2	0.058	0.08	0.72	0.47206
groupexposure:block2	0.146	0.1	1.49	0.13746
conditionunrelated:groupexposure:block2	-0.193	0.12	-1.64	0.10082
conditionrelated:groupexposure:block2	-0.061	0.12	-0.52	0.60121

Table B.6: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 4 reaction time model

	Chi-squared	Deg. Freedom	p-value
(Intercept)	3.23	1	0.0723
condition	22.259	2	1.5e-05
group	1.789	1	0.1810
block	4.291	1	0.0383
condition:group	3.997	2	0.1355
condition:block	13.574	2	0.0011
group:block	0.223	1	0.6368
condition:group:block	2.81	2	0.2453

Appendix C

Chapter 4 materials

C.1 Experiment 5

Table C.1: Model estimates for Experiment 5 logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * speaker + (1 + speaker + center(step) + /θ/ word position | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	0.93	0.21	4.4	1.1e-05
center(step)	-2.267	0.16	-14.04	< 2e-16
grouptraining	0.194	0.29	0.67	0.5010
word_posword_initial	-0.283	0.25	-1.12	0.2615
speakerfemale1	-1.011	0.22	-4.62	3.8e-06
speakermale2	-0.88	0.21	-4.26	2.0e-05
center(step):grouptraining	0.443	0.2	2.21	0.0269
center(step):word_posword_initial	-0.16	0.2	-0.79	0.4286
grouptraining:word_posword_initial	0.124	0.35	0.35	0.7239
center(step):speakerfemale1	0.214	0.18	1.17	0.2438
center(step):speakermale2	0.536	0.17	3.18	0.0015
grouptraining:speakerfemale1	0.034	0.3	0.11	0.9090
grouptraining:speakermale2	-0.218	0.28	-0.78	0.4331
word_posword_initial:speakerfemale1	1.296	0.26	4.92	8.5e-07
word_posword_initial:speakermale2	-0.608	0.26	-2.36	0.0184
center(step):grouptraining:word_posword_initial	-0.61	0.27	-2.23	0.0259
center(step):grouptraining:speakerfemale1	-0.352	0.23	-1.54	0.1247
center(step):grouptraining:speakermale2	-0.302	0.21	-1.46	0.1452
center(step):word_posword_initial:speakerfemale1	0.56	0.25	2.27	0.0234
center(step):word_posword_initial:speakermale2	0.384	0.23	1.67	0.0954
grouptraining:word_posword_initial:speakerfemale1	-0.391	0.37	-1.06	0.2869
grouptraining:word_posword_initial:speakermale2	0.364	0.35	1.03	0.3020
center(step):grouptraining:word_posword_initial:speakerfemale1	0.092	0.34	0.27	0.7873
center(step):grouptraining:word_posword_initial:speakermale2	0.567	0.31	1.82	0.0694

Table C.2: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 5 logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	19.324	1	1.1e-05
center(step)	197.261	1	< 2e-16
group	0.453	1	0.5010
word_pos	1.261	1	0.2615
speaker	25.865	2	2.4e-06
center(step):group	4.9	1	0.0269
center(step):word_pos	0.627	1	0.4286
group:word_pos	0.125	1	0.7239
center(step):speaker	11.226	2	0.0036
group:speaker	0.993	2	0.6087
word_pos:speaker	65.317	2	6.6e-15
center(step):group:word_pos	4.962	1	0.0259
center(step):group:speaker	2.833	2	0.2425
center(step):word_pos:speaker	5.167	2	0.0755
group:word_pos:speaker	5.261	2	0.0720
center(step):group:word_pos:speaker	4.995	2	0.0823

C.2 Experiment 6a

Table C.3: Model estimates for Experiment 6a logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * block + (1 + center(step) + /θ/ word position + block | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	-0.259	0.24	-1.08	0.2789
center(step)	-2.469	0.19	-12.87	<2e-16
grouptraining	0.855	0.34	2.54	0.0112
word_posword_initial	0.247	0.27	0.9	0.3670
block2	0.55	0.22	2.51	0.0120
center(step):grouptraining	0.589	0.25	2.35	0.0187
center(step):word_posword_initial	-0.253	0.18	-1.43	0.1537
grouptraining:word_posword_initial	-0.308	0.39	-0.79	0.4270
center(step):block2	0.315	0.15	2.05	0.0402
grouptraining:block2	-0.379	0.31	-1.23	0.2201
word_posword_initial:block2	0.434	0.27	1.59	0.1117
center(step):grouptraining:word_posword_initial	-0.318	0.23	-1.36	0.1753
center(step):grouptraining:block2	-0.53	0.2	-2.59	0.0095
center(step):word_posword_initial:block2	-0.708	0.26	-2.7	0.0068
grouptraining:word_posword_initial:block2	0.252	0.39	0.64	0.5209
center(step):grouptraining:word_posword_initial:block2	0.564	0.36	1.56	0.1194

Table C.4: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 6a logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	1.172	1	0.2789
center(step)	165.536	1	<2e-16
group	6.434	1	0.0112
word_pos	0.814	1	0.3670
block	6.315	1	0.0120
center(step):group	5.532	1	0.0187
center(step):word_pos	2.035	1	0.1537
group:word_pos	0.631	1	0.4270
center(step):block	4.208	1	0.0402
group:block	1.504	1	0.2201
word_pos:block	2.53	1	0.1117
center(step):group:word_pos	1.837	1	0.1753
center(step):group:block	6.733	1	0.0095
center(step):word_pos:block	7.313	1	0.0068
group:word_pos:block	0.412	1	0.5209
center(step):group:word_pos:block	2.426	1	0.1194

C.3 Experiment 6b

Table C.5: Model estimates for Experiment 6b logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * block + (1 + center(step) + /θ/ word position + block | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	-0.784	0.23	-3.48	0.00051
center(step)	-1.698	0.13	-13.16	< 2e-16
grouptraining	0.565	0.31	1.81	0.06999
word_posword_initial	-1.161	0.27	-4.34	1.5e-05
block2	0.747	0.25	2.96	0.00306
center(step):grouptraining	0.006	0.18	0.03	0.97406
center(step):word_posword_initial	0.331	0.12	2.72	0.00647
grouptraining:word_posword_initial	0.047	0.37	0.13	0.89769
center(step):block2	-0.067	0.13	-0.51	0.61185
grouptraining:block2	-0.299	0.35	-0.85	0.39334
word_posword_initial:block2	0.76	0.27	2.86	0.00427
center(step):grouptraining:word_posword_initial	-0.262	0.17	-1.55	0.12141
center(step):grouptraining:block2	-0.047	0.18	-0.26	0.79614
center(step):word_posword_initial:block2	-0.656	0.2	-3.34	0.00085
grouptraining:word_posword_initial:block2	0.199	0.37	0.54	0.58617
center(step):grouptraining:word_posword_initial:block2	0.131	0.28	0.46	0.64218

Table C.6: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 6b logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	12.093	1	0.00051
center(step)	173.189	1	< 2e-16
group	3.283	1	0.06999
word_pos	18.8	1	1.5e-05
block	8.771	1	0.00306
center(step):group	0.001	1	0.97406
center(step):word_pos	7.414	1	0.00647
group:word_pos	0.017	1	0.89769
center(step):block	0.257	1	0.61185
group:block	0.729	1	0.39334
word_pos:block	8.165	1	0.00427
center(step):group:word_pos	2.399	1	0.12141
center(step):group:block	0.067	1	0.79614
center(step):word_pos:block	11.137	1	0.00085
group:word_pos:block	0.296	1	0.58617
center(step):group:word_pos:block	0.216	1	0.64218

C.4 Experiment 6c

Table C.7: Model estimates for Experiment 6c logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * block + (1 + center(step) + /θ/ word position | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	-0.716	0.19	-3.84	0.00012
center(step)	-2.281	0.16	-14.09	< 2e-16
grouptraining	0.127	0.26	0.48	0.63111
word_posword_initial	-0.241	0.34	-0.7	0.48289
block2	1.357	0.31	4.35	1.4e-05
center(step):grouptraining	-0.124	0.22	-0.56	0.57469
center(step):word_posword_initial	0.017	0.26	0.07	0.94805
grouptraining:word_posword_initial	0.543	0.47	1.16	0.24462
center(step):block2	-0.455	0.3	-1.5	0.13427
grouptraining:block2	-0.915	0.44	-2.06	0.03919
word_posword_initial:block2	-0.262	0.45	-0.59	0.55749
center(step):grouptraining:word_posword_initial	0.4	0.36	1.12	0.26170
center(step):grouptraining:block2	0.198	0.45	0.44	0.66265
center(step):word_posword_initial:block2	0.042	0.43	0.1	0.92327
grouptraining:word_posword_initial:block2	0.633	0.62	1.02	0.30632
center(step):grouptraining:word_posword_initial:block2	0.003	0.62	0.01	0.99593

Table C.8: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 6c logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	14.756	1	0.00012
center(step)	198.425	1	< 2e-16
group	0.231	1	0.63111
word_pos	0.492	1	0.48289
block	18.907	1	1.4e-05
center(step):group	0.315	1	0.57469
center(step):word_pos	0.004	1	0.94805
group:word_pos	1.354	1	0.24462
center(step):block	2.242	1	0.13427
group:block	4.252	1	0.03919
word_pos:block	0.344	1	0.55749
center(step):group:word_pos	1.26	1	0.26170
center(step):group:block	0.19	1	0.66265
center(step):word_pos:block	0.009	1	0.92327
group:word_pos:block	1.046	1	0.30632
center(step):group:word_pos:block	0	1	0.99593

C.5 Experiment 7a

Table C.9: Model estimates for Experiment 7a logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * block + (1 + center(step) + /θ/ word position + block | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	0.766	0.26	2.92	0.0035
center(step)	-2.609	0.2	-13.31	< 2e-16
grouptraining	0.58	0.38	1.54	0.1231
word_posword_initial	0.524	0.31	1.67	0.0959
block2	1.804	0.32	5.58	2.4e-08
center(step):grouptraining	-0.182	0.27	-0.67	0.4999
center(step):word_posword_initial	-0.362	0.28	-1.3	0.1924
grouptraining:word_posword_initial	-0.44	0.48	-0.92	0.3561
center(step):block2	0.076	0.22	0.35	0.7256
grouptraining:block2	-0.249	0.49	-0.51	0.6076
word_posword_initial:block2	-1.171	0.4	-2.96	0.0031
center(step):grouptraining:word_posword_initial	-0.809	0.5	-1.61	0.1067
center(step):grouptraining:block2	-0.097	0.34	-0.29	0.7717
center(step):word_posword_initial:block2	-0.217	0.38	-0.57	0.5697
grouptraining:word_posword_initial:block2	-0.472	0.6	-0.79	0.4314
center(step):grouptraining:word_posword_initial:block2	1.554	0.63	2.45	0.0143

Table C.10: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 7a logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	8.518	1	0.0035
center(step)	177.033	1	< 2e-16
group	2.377	1	0.1231
word_pos	2.772	1	0.0959
block	31.101	1	2.4e-08
center(step):group	0.455	1	0.4999
center(step):word_pos	1.699	1	0.1924
group:word_pos	0.852	1	0.3561
center(step):block	0.123	1	0.7256
group:block	0.264	1	0.6076
word_pos:block	8.77	1	0.0031
center(step):group:word_pos	2.602	1	0.1067
center(step):group:block	0.084	1	0.7717
center(step):word_pos:block	0.323	1	0.5697
group:word_pos:block	0.619	1	0.4314
center(step):group:word_pos:block	5.997	1	0.0143

C.6 Experiment 7b

Table C.11: Model estimates for Experiment 7b logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * block + (1 + center(step) + /θ/ word position | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	1.606	0.25	6.49	8.3e-11
center(step)	-1.94	0.17	-11.29	< 2e-16
grouptraining	-0.9	0.35	-2.6	0.0092
word_posword_initial	-2.079	0.26	-7.95	1.8e-15
block2	-0.865	0.18	-4.7	2.6e-06
center(step):grouptraining	-0.364	0.25	-1.47	0.1410
center(step):word_posword_initial	-0.328	0.14	-2.39	0.0167
grouptraining:word_posword_initial	0.11	0.39	0.28	0.7795
center(step):block2	-0.069	0.12	-0.58	0.5601
grouptraining:block2	0.853	0.26	3.23	0.0012
word_posword_initial:block2	1.055	0.27	3.96	7.6e-05
center(step):grouptraining:word_posword_initial	-0.975	0.31	-3.13	0.0018
center(step):grouptraining:block2	0.105	0.19	0.55	0.5810
center(step):word_posword_initial:block2	-0.533	0.21	-2.51	0.0122
grouptraining:word_posword_initial:block2	-0.407	0.41	-0.98	0.3248
center(step):grouptraining:word_posword_initial:block2	0.938	0.4	2.32	0.0206

Table C.12: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 7b logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	42.183	1	8.3e-11
center(step)	127.518	1	$<$ 2e-16
group	6.784	1	0.0092
word_pos	63.241	1	1.8e-15
block	22.06	1	2.6e-06
center(step):group	2.167	1	0.1410
center(step):word_pos	5.726	1	0.0167
group:word_pos	0.078	1	0.7795
center(step):block	0.34	1	0.5601
group:block	10.431	1	0.0012
word_pos:block	15.658	1	7.6e-05
center(step):group:word_pos	9.769	1	0.0018
center(step):group:block	0.305	1	0.5810
center(step):word_pos:block	6.276	1	0.0122
group:word_pos:block	0.97	1	0.3248
center(step):group:word_pos:block	5.363	1	0.0206

C.7 Experiment 7c

Table C.13: Model estimates for Experiment 7c logistic regression analysis. Generalized linear mixed model fit by maximum likelihood (Laplace Approximation). Family:binomial (logit). Formula: response = /θ/ ∼ center(step) * group * /θ/ word position * block + (1 + center(step) + /θ/ word position + block | subj).

	Coeff.	Std. Error	z-value	p-value
(Intercept)	1.334	0.3	4.5	6.9e-06
center(step)	-2.683	0.25	-10.72	< 2e-16
grouptraining	0.967	0.42	2.28	0.0225
word_posword_initial	0.407	0.33	1.23	0.2168
block2	0.789	0.27	2.94	0.0033
center(step):grouptraining	-0.365	0.35	-1.06	0.2904
center(step):word_posword_initial	-0.607	0.26	-2.31	0.0206
grouptraining:word_posword_initial	-0.217	0.51	-0.43	0.6680
center(step):block2	-0.317	0.23	-1.41	0.1585
grouptraining:block2	-0.707	0.4	-1.75	0.0799
word_posword_initial:block2	-0.475	0.42	-1.12	0.2618
center(step):grouptraining:word_posword_initial	-0.592	0.46	-1.29	0.1981
center(step):grouptraining:block2	0.164	0.35	0.47	0.6392
center(step):word_posword_initial:block2	0.207	0.4	0.51	0.6082
grouptraining:word_posword_initial:block2	0.786	0.66	1.2	0.2303
center(step):grouptraining:word_posword_initial:block2	0.335	0.66	0.51	0.6104

Table C.14: Analysis of Deviance Table (Type III Wald chisquare tests) for Experiment 7c logistic regression model.

	Chi-squared	Deg. Freedom	p-value
(Intercept)	20.225	1	6.9e-06
center(step)	115.009	1	< 2e-16
group	5.206	1	0.0225
word_pos	1.525	1	0.2168
block	8.661	1	0.0033
center(step):group	1.118	1	0.2904
center(step):word_pos	5.358	1	0.0206
group:word_pos	0.184	1	0.6680
center(step):block	1.988	1	0.1585
group:block	3.066	1	0.0799
word_pos:block	1.259	1	0.2618
center(step):group:word_pos	1.656	1	0.1981
center(step):group:block	0.22	1	0.6392
center(step):word_pos:block	0.263	1	0.6082
group:word_pos:block	1.439	1	0.2303
center(step):group:word_pos:block	0.26	1	0.6104