8-24-2015

Phonetic Adaptation to Foreign-Accented Speech

Xin Xie
University of Connecticut - Storrs, xin.xie@uconn.edu

Phonetic Adaptation to Foreign-Accented Speech

Xin Xie, PhD

University of Connecticut, 2015

Over the past few decades, there has been considerable effort to find the mechanisms through which adult listeners can accommodate the rampant phonetic variation in natural speech. My dissertation concerns one source of variability: phonetic variation in speech produced by individuals with foreign accents. Mounting evidence shows that listeners not only adapt to specific speakers by adjusting acoustic-phonetic mappings, they also sometimes generalize the remapping to novel talkers. In this dissertation, I present a series of experiments examining the mechanism of rapid phonetic adaptation and its generalization across talkers. I tested native-English listeners' adaptation to Mandarin-accented English words, focusing on /d/ in word-final position. The first set of experiments (Experiments 1-3) investigated talker-specific adaptation. I found that perceptual learning for speech was not just a matter of adjusting phonetic boundaries in face of noncanonical tokens; it also promoted a reorganization of the internal category structure. The learning resulted in changes in cue-weighting functions that may prepare listeners for adapting to similar variation in other acoustic environments. The second set of experiments (Experiments 4, 5A and 5B) examined generalization of learning across talkers following *single*-talker exposure or *multiple*-talker exposure. *Single*-talker exposure failed to produce generalization to a novel talker. Following *multiple*-talker exposure, cross-talker generalization was evident only when the test talker (a novel talker) was acoustically similar to (one or more of) the exposure talkers. Lastly, Experiments 6 and 7 present case studies of talker-specific adaptation to foreign-accented speakers, showing a role of speaker intelligibility and within-

talker variability in phonetic adaptation. In summary, the results of these experiments demonstrate that the lexically-guided phonetic reorganization mechanism that substantiates the adaptation to idiolect differences of native speakers also supports adaptation to natural foreign accents. In addition, bottom-up similarity at the acoustic-phonetic level explains a range of situations in which adaptation effects may or may not generalize to novel talkers. Taken together, the findings advance our understanding of the reorganization of the perceptual architecture that listeners experience when they adjust to unfamiliar speech.

Phonetic Adaptation to Foreign-Accented Speech

by

Xin Xie

B.S., Zhejiang University, 2009

M.A., University of Connecticut, 2014

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Phonetic Adaptation to Foreign-Accented Speech

Presented by

Xin Xie, B.S., M.A.

Major Advisor

_____

Emily B. Myers

Co-Major Advisor

_____

Carol A. Fowler

Associate Advisor

_____

Rachel M. Theodore

Associate Advisor

_____

James S. Magnuson

University of Connecticut
2015

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to many people who walked with me on this journey. Without your help and support, this dissertation would not be possible. First, my deepest gratitude goes to my committee members. I owe everything to my major advisor, Dr. Emily Myers. Your expertise, devotion, patience and trust in me have led to the completion of this project. You have been a great mentor and a role model that I look up to in my academic career. Dr. Carol Fowler, you are the reason that I entered University of Connecticut in the first place. Your advice and supervision of this work are invaluable to me. Thank you for your precious insights and your prompt responses to all my questions throughout the years. Dr. Rachel Theodore, you have brought in a unique perspective to this project. I am grateful for extremely helpful discussions on research and work-life balance. Thank you for guiding and sharing. Dr. James Magnuson, your insightful suggestions on the experiment design and valuable comments on the draft of this dissertation have helped to put this work into a much better shape. I learned a lot.

My gratitude also goes to other faculty members who have provided insights and support throughout the project. Thank you, Dr. Gerry Altmann, Dr. Len Katz and Dr. Jay Rueckl for being my reviewers and for providing critical comments. I am grateful to Dr. Adam Sheya, who kindly offered consultation on running mixed-effects models in R. I would also like to thank my dear and wonderful labmates and friends in the Department of Psychology, who have showed unconditional support. I enjoy your comradeship! For administrative support, I am thankful to Debbie Vardon and Carol Valone.

In the real world, my gratitude to my family and my friends is inexpressible. Special thanks to Wen, Lisi, Lei and Yingying. We've spent the most time together under the same roof.

You made me laugh and strive. Lastly, to whom I dedicate this work, my family. My parents, thank you for countless days of unconditional love and devotion. Your encouragement instilled determination in me across the Pacific. And Jiesi, you are the last piece of the jigsaw puzzle of my life, but the first person to share my joy and frustration. Thank you for making the journey easier and more fun. And lastly of course, thanks for being a pilot speaker in my experiments and providing the recordings (sorry that you were not chosen as the test speaker).

TABLE OF CONTENTS

CHAPTER 1 GENERAL INTRODUCTION

Natural speech exhibits substantial acoustic-phonetic variation such that, as speech context varies, different acoustic patterns may denote the same linguistic information, leading to the "Lack of Invariance" problem (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Primary among many sources of variability is talker differences. Each speaker represents a unique combination of age, gender, vocal tract anatomy, idiosyncratic speaking style, and long-term language experience (e.g., regional dialect, native or non-native, bilingual or monolingual). Talker variability is manifested as a very wide variety of audible acoustic-phonetic variation in speech production, which further leads to differences in perceptual tasks (e.g., Peterson & Barney, 1952; Strand & Johnson, 1996; Allen & Miller, 2004; Theodore, Miller, & DeSteno, 2009). Despite this variation, listeners readily identify spoken words across novel talkers.

Extensive research has investigated the perceptual operations that underlie listeners' ability to accommodate talker variability in speech perception. Early studies focused on how listeners resolve talker differences in perceiving typical native speech. A persistent debate originating from this body of research concerns the role of talker information in linguistic representations. Abstractionist approaches assume that lexical entries take abstract forms in the mental lexicon and word recognition is mediated by pre-lexical representations (McClelland & Elman, 1986; Norris, McQueen, & Cutler, 2000; Marslen-Wilson & Warren, 1994; Gaskell & Marslen-Wilson, 2002). Speech perception involves mapping a set of acoustic properties onto canonical representations of linguistic categories; information about individual talkers is not intrinsic to the abstract representations. Talker variability is considered a kind of noise that is eliminated by normalization processes which rescale acoustic parameters (Mann & Repp, 1980; Nearey, 1989; McGowan & Cushing, 1999). At another extreme, episodic approaches

(Goldinger, 1996, 1998; Johnson, 1997; 2006) postulate no abstract pre-lexical representations. Rather, talker detail of spoken words is integrally stored as part of a word's representation in memory. Empirical evidence has challenged both approaches. Most notably, problematic for abstractionist theories, talker characteristics are used in lexical retrieval (Nygaard, Sommers, & Pisoni, 1995; Bradlow, Nygaard, & Pisoni, 1999). Meanwhile, the human speech perceptual system is more robust than predicted by episodic theories: listeners readily generalize what they learn from specific spoken instances to novel words (Nygaard, Sommers, & Pisoni, 1994), novel phonetic contexts (Theodore & Miller, 2010), and in some cases, novel talkers (Bradlow & Bent, 2008).

The recent literature on "perceptual learning for speech" sheds new light on talker accommodation by highlighting the plasticity of perceptual processes: as native listeners encounter unfamiliar pronunciations that cause perceptual ambiguity, they use top-down information (e.g., lexical knowledge or visual information) to constrain the interpretation of the ambiguous sound and alter the sound-to-category mapping accordingly (Norris, McQueen, & Cutler, 2003). For example, if listeners hear a speaker pronouncing a sound ambiguous between /s/ and /f/ (denoted here as /?/), then hearing the sound in a carrier word such as 'belie?' ('belief') would bias the interpretation of it as /f/. This exposure also affects subsequent interpretation of other similar ambiguous sounds in a way consistent with prior exposure. This kind of lexically-guided phonetic retuning requires very brief exposure (as few as ten critical items, see Kraljic & Samuel, 2007). Further, the altered mapping is maintained for a given talker (e.g., Kraljic & Samuel, 2005; Eisner & McQueen, 2006) and readily generalizes across the lexicon (e.g., McQueen, Cutler, & Norris, 2006; Sjerps & McQueen, 2010).

Beyond this empirical literature on adaptation to *artificially-altered native* phonemic productions, other research has examined adaptation to *natural foreign-accented speech*, which is often conceived as an extreme case of 'unfamiliar pronunciation'. Non-native accented-speech can be substantially different from native speech. This deviation is manifested as differences in the acoustic distributions of speech tokens along multiple dimensions for multiple categories (e.g., Flege, MacKay, & Meador, 1999; Flege, Munro, & Skelton, 1992), making recognition of non-native speech effortful and often times, inaccurate (e.g., Munro & Derwing, 1995). For instance, listeners might have to overcome multiple acoustic mismatches to correctly hear a 'thick pad' instead of a 'sick pet' in an unfamiliar foreign accent. A number of studies on second language (L2) speech intelligibility show that as listeners gain experience with a foreign-accented speaker, recognition of words and sentences produced by that speaker becomes more accurate (Gass & Varonis, 1984; Weil, 2001; Clarke & Garrett, 2004). This line of research has explored a range of situations in which listeners adapt to speakers of foreign accents and generalize across talkers (Bradlow & Bent, 2008; Wade, Jongman, & Sereno, 2007). However, since accuracy in sentence or word transcription tasks is taken as the measure of adaptation in these studies, much less is known about perceptual or representational changes that occur at the phonemic or subphonemic level.

Emerging evidence has begun to show that a mechanism of lexically-guided phonetic retuning, which helps listeners with idiosyncratic pronunciations in native speech, also supports adaptation to foreign-accented speakers (e.g., Sumner, 2011; Eisner, Melinger, & Weber, 2013). The findings on recalibration of phonetic categories as a way to accommodate talker-specific pronunciations (in native- and foreign-accented speech) add a new twist to the debate on the "degree of abstraction" in lexical representations. On the one hand, the adaptation happens pre-

lexically: improvement in word recognition does not require exposure to specific spoken instances; rather, learning is more rapidly generalized to novel words than predicted by episodic theories. On the other hand, a number of studies suggest that learning does not necessarily transfer to new talkers (e.g., Kraljic & Samuel, 2005; Eisner & McQueen, 2005). This suggests that listeners are able to maintain multiple pre-lexical representations (or multiple talker-specific acoustic-to-phoneme mapping algorithms), either at the phonemic level or sub-phonemic level. A single canonical representation as conceived in abstractionist theories apparently does not account for the data.

Given the growing evidence on perceptual adaptation, the right question to ask is perhaps not whether linguistic representations are abstract or episodic, but rather, how more and less abstract representations coexist and collectively affect speech processing. Perceptual reorganization of phonetic categories offers a special window for examining how listeners may incorporate novel instances into existing representations and potentially form new abstract representations as they gain more experience with previously unfamiliar pronunciations. A number of important questions about the dynamic adaptation processes remain unanswered. *First, what is the scope of phonetic reorganization?* Much of existing research on phonetic adaptation has focused on changes in category boundary locations as listeners categorize tokens from two phonetic categories. It is an open question whether perceptual learning produces a pervasive reorganization of the phonetic category structure beyond the boundary region. *Second, what is the mechanism by which an altered sound-to-category mapping generalizes to novel talkers?* There is insufficient research into cross-talker generalization of phonetic adaptation to draw any firm conclusions. Of note, the aforementioned two lines of research (studies of phonetic recalibration of specific categories and foreign accent intelligibility studies) have both explored

various conditions for talker generalization (e.g., Kraljic & Samuel, 2007; Bradlow & Bent, 2008). However, the use of different paradigms makes it difficult to interpret discrepant generalization patterns. *Third, what are the limits of phonetic reorganization, specifically what type of speech input is required for rapid phonetic adaptation?* Notably, only very brief exposure is needed for native listeners to adapt to idiosyncratic *native* speakers, in contrast with reports of more effortful adaptation to *foreign-accented* speech, which usually takes multiple training sessions (Wade et al., 2007; Bradlow & Bent, 2008). Are there speaker-related factors that slow down adaptation? Answers to these questions are of significance for theoretical advances on the nature of sub-lexical representations; they are also important for a practical understanding of speech plasticity.

My dissertation addresses these three questions via a series of experiments examining how phonetic adaptation and generalization to novel talkers might occur in the context of rich within- and inter-talker variability in a foreign accent. These experiments are intended to facilitate a unification of findings originating from different lines of research and, in this way, provide an integrated account of how listeners represent and adapt to unfamiliar pronunciations in speech. Chapter 2 reviews existing empirical work on talker-specific adaptation and generalization across talkers, focusing on work that investigates phonetic recalibration of specific segments in native-accented speech (e.g., Norris et al., 2003; Kraljic & Samuel, 2005, 2006, 2007) and studies of intelligibility in non-native accents that emphasize natural variability (e.g., Bradlow & Bent, 2008). I then identify a few remaining questions, discuss why answering them has important theoretical implications, and briefly introduce the set of experiments designed to address each question. Chapter 3-5 presents the experiments in detail. Chapter 3 presents experiments exploring listeners' adaptation to phonetic variation in foreign-accented words,

asking whether the natural variation in these tokens produces pervasive reorganization of the phonetic category structure, and whether this information is used to map to the lexicon. Chapter 4 details experiments on cross-talker generalization of accent learning in two exposure conditions: single-talker exposure and multiple-talker exposure, aiming to pinpoint the mechanism that subserves talker generalization of phonetic reorganization. Chapter 5 further presents two case studies of talker-specific adaptation to foreign-accented speakers, investigating the role of speaker intelligibility and within-talker variability in phonetic adaptation. Finally, Chapter 6 presents a general discussion of findings from Chapters 3, 4 and 5 and their implications for theories of speech perception and adaptation.

CHAPTER 2 PERCEPTUAL ADAPTATION TO ACCOMMODATE TALKER VARIATION

It has been long observed that speakers demonstrate distinctive acoustic patterns in the realizations of phonemes and words (e.g., Peterson & Barney, 1952; Dorman, Studdert-Kennedy, & Raphael, 1977; Newman, Clouse, & Burnham, 2001), making the mapping from sound signal to linguistic representations a talker-contingent process (e.g., Johnson, 1990; Nygaard et al., 1994; Allen & Miller, 2004; Theodore & Miller, 2010). Theories debate the nature of stored representations, whether they are *talker-specific* (episodic theories: Goldinger, 1996, 1998; John, 1997; 2006) or abstract and *talker-independent* (abstract theories: e.g., McClelland & Elman, 1986; Norris & McQueen, 2008; Norris et al., 2000).

Despite the fundamental differences between the abstract and episodic theories, both approaches assume that speech recognition is a pattern matching process that passively maps sounds onto existing representations, either via a single fixed mapping, or via potentially unlimited numbers of mappings. Nusbaum and colleagues (Nusbaum & Morin, 1992; Nusbaum & Henly, 1992; Nusbaum & Magnuson, 1997; Magnuson & Nusbaum, 2007) proposed a *contextual tuning* theory which views speech recognition as an adaptive process: listeners *tune* acoustic-phonetic mappings using a talker's acoustic-phonetic space as the *context for tuning.* Two important aspects of this theory are noteworthy: first, it acknowledges that the mapping from acoustic cues to categories (phonemes or words) is *nondeterministic*. For instance, individual talkers may use different cues to denote phonemes (e.g., Dorman et al., 1977). Support for the theory comes from evidence showing that listeners selectively attend to different cues in the presence of talker variability (i.e., a single talker vs. multiple talkers, Johnson, 1991; Nusbaum & Morin, 1992; Wong, Nusbaum, & Small, 2004). Second, it suggests that talker tuning is a time-consuming and cognitively demanding process and is modulated by listeners'

attention and expectations. Recognition of phonemes and syllables is slower or more errorful as listeners encounter a new talker or have to switch between talkers (Kakehi, 1992; Nusbaum & Morin, 1992). When there is minimal talker variability, the switching cost between talkers is modulated by listeners' expectation of talker changes (Magnuson & Nusbaum, 2007). It is suggested that context tuning occurs only in cases where talker changes result in perceptual uncertainty. In other words, listeners do not have to retune the acoustic-phonetic mapping for each talker and once retuned, the mapping is retained till further talker changes require another round of tuning. In order to account for the findings that talker-specific information has long-term effects (e.g., Nygaard et al., 1994), the authors suggested that talker detail may be stored separately from abstract representations of linguistic categories in memory. One shortcoming of this theory is the lack of clarity in specifying how perceptual processes could be tailored to accommodate talker-related phonetic variation. Recent research on perceptual learning for speech takes an important step toward finding a clear mechanism of how talker-specific adaptation might be achieved. Moreover, it poses problems for the context tuning theory in that results of talker-specific adaptation are maintained and applied to contextually impoverished instances (e.g., nonwords).

Talker-Specific Adaptation: What do Listeners Adapt to?

Accommodating Idiosyncratic Pronunciations in Native-Accented Speech

The literature on perceptual learning of native phoneme contrasts shows that listeners recalibrate phonetic boundaries under the guidance of top-down knowledge (e.g., Norris et al., 2003; Kraljic & Samuel, 2005, 2006, 2007). In a typical version of perceptual learning for speech paradigm, native listeners hear artificially-created ambiguous tokens (e.g., midway between /s/ and /f/) presented in lexically-biased contexts. For example, the ambiguous sound

(“?”) is embedded either in /f/-biased words (e.g., *belie?*) or, for different participants, in /s/-biased words (e.g., *Pari?*). The critical manipulation of lexically-biased contexts enables listeners to resolve acoustic ambiguity using their lexical knowledge. After this initial exposure, listeners then identify consonant tokens along an acoustic continuum (e.g., /ɛf/ -/ɛs/). The perceptual learning is manifested as a between-group difference in the phonetic boundary in a direction specified by the lexical context during exposure such that the ambiguous sound is incorporated into the recalibrated phonetic category. Similar findings are replicated in perceptual learning for stop consonants (Kraljic & Samuel, 2006, 2007) and vowels (Maye, Aslin, & Tanenhaus, 2008). Segmental adjustments linked to a specific talker facilitate subsequent recognition of spoken words, generalizing to untrained words containing the critical segment (McQueen et al., 2006; Sjerps & McQueen, 2010), to the same segment across word positions (Jesse & McQueen, 2011) and to other segments across place of articulation (for stops, Kraljic & Samuel, 2006). Further, once adjusted, the new acoustic-to-phoneme mapping is maintained for a given speaker (e.g., Kraljic & Samuel, 2005; Eisner & McQueen, 2006).

Accommodating Non-Native Talker Variation in Foreign-Accented Speech

In general, these phonetic adjustments are taken to reflect the mechanism by which listeners handle acoustic signals that deviate from canonical pronunciations, such as idiosyncratic pronunciations or foreign accents (see Samuel, 2011 for a review). A number of studies have adapted the perceptual learning paradigm (Norris et al., 2003) and show that phonetic retuning contributes to accent adaptation, at least in part. When acoustic-phonetic variation of accented words mismatches existing words, listeners use lexical knowledge to retune the mapping from the acoustic signal to native phonetic categories (Sumner, 2011; Reinisch & Holt, 2014). For instance, a French-accented /p/ sound is acoustically more similar to a /b/ (than a /p/) when

mapped onto distributions of native English. However, when French-accented /p/ tokens are heard in real English-words like "*paint*", native-English listeners recalibrate the category boundary between /b/ and /p/ (Sumner, 2011). Thus, listeners show great flexibility in adaptation to foreign-accented speech, despite its considerable acoustic-phonetic deviation from native norms.

Importantly, phonetic remapping further facilitates word recognition in foreign-accented speech (e.g., Eisner, Melinger, & Weber, 2013; Witteman, Weber, & McQueen, 2013). Eisner et al. (2013) examined the process by which native English listeners adapted to Dutch-accented English in which, as in Dutch itself, final obstruents were devoiced. For instance, the devoicing rule will make word-final /d/s acoustically and perceptually similar to members of the unintended category /t/ (Warner, Jongman, Sereno, & Kemps, 2004) such that a word like 'seed' will be produced similar to the word 'seat.' In this study, exposure to multisyllabic words (e.g., *overload*) containing devoiced /d/ in word-final position produced changes in the priming of /d/-final words. Specifically, the accented production of *seed,* sounding like [si:t^{h}], primed the written form, "SEED" to a greater extent in listeners who had heard /d/-final words during exposure than in listeners who did not have this exposure (no significant identity *seed* ([si:t^{h}])-SEED priming for control listeners). This suggests that listeners accepted a Dutch-accented /d/ as a production of /d/ category following exposure.

Remaining Questions about Talker-Specific Adaptation

Despite ample evidence that phonetic retuning bolsters rapid adaptation to unfamiliar pronunciations (native and foreign-accented), there are a few reasons why existing research has not provided a complete picture of how listeners adapt to unfamiliar pronunciations. One limitation is that investigations have almost exclusively measured phonetic retuning in terms of a

recalibration of phonetic category boundaries (Norris et al., 2003; Kraljic & Samuel, 2005, 2006, 2007; Reinisch & Holt, 2014). It is well established that phonetic categories have a graded internal structure such that some members of the category may be represented as better exemplars than others, as revealed by overt judgment and covert psychological responses (e.g., Miller & Volaitis, 1989; Kuhl, 1991; Samuel, 1982). This sensitivity to sub-phonemic variation cascades to lexical processing and has a gradient effect on lexical activation (Warren & Marslen-Wilson, 1987, 1988; Marslen-Wilson, Moss, & van Halen, 1996). A wide body of research suggests that the "goodness of fit" between incoming speech and lexical representations influences the activation of a lexical entry and its acoustic-phonetic competitors (Andruski, Blumstein, & Burton, 1994; Utman, Blumstein & Burton, 2000).

Could a reorganization of internal category structure drive the improved recognition of foreign-accented words without a shift of phonetic category boundary? Of note, naturally-produced foreign-accented tokens are not always as ambiguous as categories are in the research on experimentally-controlled speech sounds (e.g., Norris et al., 2003; Reinisch & Holt, 2014). Instead, they may have well-defined category membership but still exhibit salient acoustic-phonetic deviation from native speech. For example, the Spanish vowel /u/ tends to have a lower F2 frequency than English /u/ (Bradlow, 1995). Despite this variation, productions of this particular non-native phoneme rarely cause confusion regarding phoneme identity when perceived by native-English listeners (e.g., Wade et al., 2007). Moreover, highly intelligible non-native speakers may produce clear (in terms of phoneme identity), albeit atypical, speech tokens. Some adaptation may be required to process these kinds of deviations efficiently. Given the gradient effect of acoustic-phonetic variation on lexical activation, even adaptation within a category may substantially improve word recognition.

Early work demonstrates that the representation of phonetic structure is malleable. Listeners readily adjust both phonetic boundaries and best exemplars of a phonetic category (Miller & Volaitis, 1989; Volaitis & Miller, 1992) in the face of acoustic variation arising from contextual variables, such as speaking rate or place of articulation. In addition, while some contextual factors have effects on both boundary location and internal category structure, lexical status, for instance, affects only the location of between-category boundaries (Ganong, 1980) but does not change the location of best-exemplar region within a category (Allen & Miller, 2001). Such a dissociation marks the importance of considering both the internal structure of categories and phonetic boundaries in understanding the plasticity of phonetic representation. In addition, listeners are known to track sub-phonemic detail in a talker-specific manner, for example, linking talker identity with talkers' tendency to produce word-initial stops with short versus long voice onset time (Allen & Miller, 2004; Theodore & Miller, 2010). In theory, this ability to store talker-specific experience can prepare listeners to use it in guiding perceptual adaptation. It is an open issue whether the lexically-guided phonetic adjustments as observed in perceptual learning of accented speakers entail changes beyond the phonetic boundary region to also change perceived goodness of tokens throughout the phonetic category.

A second limitation of research on foreign accents concerns effects of perceptual adaptation in lexical access. Foreign-accented sounds of different categories may fall into a single category to native listeners. For example, both a devoiced /d/ and a normal /t/ in Dutch-accented English sound like /t/ to English listeners, as shown in Eisner et al. (2013). What remains unclear is the effect of adaptation on lexical competition; that is, whether an accented *seed* would activate not only 'seed' (the intended target) but also 'seat' (the most surface-similar form) in the mental lexicon. It is long observed that word recognition depends not only on the

degree of fit between the speech signal and the stored representation of a lexical candidate, but also on the competition between multiple simultaneously activated lexical representations (the intended candidate and other phonetically similar competitors which partially match the sound signal, e.g., "*seat*" for "*seed*"; Marslen-Wilson, Nix, & Gaskell, 1987; Luce, Pisoni, & Goldinger, 1990). Indeed, an entirely ambiguous token (e.g., midway between /d/ and /t/) can activate both alternative interpretations of the sound significantly (e.g., Connine, Blasko, & Wang, 1994). In Eisner et al. (2013), devoiced /d/-final words primed the printed identical words (e.g., auditory "seed" – visual "SEED") to a greater extent after adaptation; however, the increase in identity priming for devoiced /d/ words in itself does not inform the extent of completeness of learning: it may be that native listeners still face quite an amount of lexical competition upon hearing accented variants of devoiced /d/ words, despite an increase of lexical activation over baseline for the intended candidate. That is, accented 'seed' may continue to robustly prime 'seat', even though listeners have learned that this pronunciation maps to 'seed'.

A few studies investigating adaptation to ambiguous sounds in native speech have found a complete elimination of lexical competitors following perceptual learning (e.g., McQueen et al., 2006; Sjerps & McQueen, 2010). Yet it is unknown whether such complete learning can be obtained for foreign-accented phonetic variation, given other evidence that adapting to a foreign accent is much harder than adapting to a native accent (compare Trude & Brown-Schmidt, 2012 to Trude, Tremblay, & Brown-Schmidt, 2013). Experiments 1-3 (presented in Chapter 3) are designed to overcome the two limitations noted above by examining a) whether perceptual adaptation to a foreign-accented speaker instigates a more pervasive effect both within and between phonetic categories than has been previously investigated, and b) whether these changes help attenuate lexical competition.

Generalization of Adaptation: When do Listeners Generalize to Novel Talkers?

Cross-Talker Generalization of Phonetic Retuning in Native-Accented Speech

As reviewed above, exposure to unfamiliar pronunciations may fundamentally change the phonetic analysis of the speech signal for a specific talker. An important question is whether listeners maintain the altered mapping just for the specific talker or whether they generalize across talkers. The answer to this question is central to the debate on the nature of pre-lexical representations. Empirical investigations of phonetic retuning for ambiguous phoneme contrasts in native-accented speech have yielded mixed results. On the one hand, listeners adapt in a talker-specific manner for fricatives. After adapting to a speaker's productions of ambiguous fricatives, they do not apply the adjusted "phonemic representation" in the perception of a different talker when tested immediately after the initial exposure (Kraljic & Samuel, 2005). An altered mapping for a specific talker is maintained over a 12-hour interval, despite intervening speech stimuli from other talkers (Eisner & McQueen, 2006). Furthermore, once a sound-to-category mapping is adjusted, hearing conflicting tokens from a different talker does not undermine previous perceptual learning results but hearing them from the same speaker does, suggesting listeners keep person-specific representations separate (Kraljic & Samuel, 2007). On the other hand, adaptation for stop consonants has been shown to be talker-independent. Kraljic and Samuel (2006) exposed listeners to a male speaker's ambiguous productions (midway between /d/ and /t/) in /d/-biased words (e.g., *kingdom*). Following this exposure, there was an increase of /d/ reports in categorizing ambiguous sounds along a nonword-nonword (e.g., /ada/-/ata/) continuum, indicating a shift of the category boundary location between /d/ and /t/. Importantly, the boundary shift was evident regardless of who was speaking (the exposure male speaker or another unfamiliar female speaker). In Kraljic and Samuel (2007), listeners adjusted

their representation for /d-/t/ phoneme contrast in the same way following exposure to a male speaker. However, adjustments were reset to baseline when listeners later heard the same ambiguous sounds embedded in /t/-biased words (e.g., *cafeteria*, which conflicted with prior experience), even though the new sounds were embedded in a female voice. Taken together, talker generalization pattern seems to differ between phoneme classes within this literature.

Cross-Talker Generalization of Adaptation to Foreign-accented Speech

Another line of research has investigated whether adaptation to a foreign-accented speaker generalizes to a novel talker of the same accent, using a paradigm very different than those for the "perceptual learning for speech" studies. In this line of work, non-native accents are assessed using intelligibility measures; reports have consistently shown that single-talker training does not enhance speech intelligibility of a different talker with the same accent (Jongman, Wade, & Sereno, 2003; Bradlow & Bent, 2008). However, exposure to multiple talkers who share a foreign accent appears to enhance intelligibility of other talkers with the same accent in some cases (Bradlow & Bent, 2008; Sidaras, Alexander, & Nygaard, 2009; but see Clarke, 2000; Wade et al., 2007 for negative evidence). Bradlow and Bent (2008) trained native-English listeners to recognize sentences in Mandarin-accented English. They found that having heard sentences produced by multiple Mandarin-accented talkers, listeners showed an improvement in recognizing untrained sentences from a novel Mandarin-accented talker. Such facilitation of sentence recognition was equivalent to facilitation effects elicited by training on the same talker that was used at test. Similar findings were reported by Sidaras et al. (2009) who trained listeners to transcribe words from a group of Spanish-accented speakers. Following multiple-talker training, listeners transcribed untrained words spoken by a new group of Spanish-accented speakers. Their performance was as good as participants who were previously trained with the

test speakers and was better than control participants who had no pre-test exposure to the particular accent.

The question is: what is afforded by multiple-talker exposure, but not by single talker exposure, that allows generalization to a novel talker? It was suggested by authors of these two studies that exposure to multiple talkers with the same accent enabled listeners to learn the acoustic-phonetic regularities in the accent which helped them to tag certain types of acoustic variability as characteristic of a language community rather than characteristic of a specific talker (see also the discussion of Baese-Berk, Bradlow, & Wright 2013). However, in these studies, adaptation has been exclusively measured by an increase of word recognition accuracy in transcription tasks, which cannot in and of itself unequivocally support the hypothesis. For instance, it is possible that the increased variability in the form of multiple talkers causes a general relaxation of the mapping from nonstandard speech tokens to word forms (since all speech tokens have to be real words in a transcription task), allowing many possible acoustic tokens to map to a phoneme, without instigating any changes in specific segmental representations (e.g., Brouwer, Mitterer, & Huettig, 2012; McQueen & Huettig, 2012). Similarly, the null effects of single talker exposure could indicate a lack of generalization of phonetic adjustments, or alternatively, it could be that the test measures of global intelligibility are not sensitive enough to detect talker-independent generalization for specific phoneme contrasts.

To our knowledge, only two studies on foreign-accented speech have examined cross-talker generalization at the phoneme level (Witteman et al., 2013; Reinisch & Holt, 2014). Reinisch and Holt (2014) adopted the paradigm of Norris et al. (2003) and examined native-English listeners' adaptation to artificially-created ambiguous sounds (midway between /s/ and /f/) embedded in Dutch-accented English. A separate /f/-/s/ continuum was constructed for each

speaker (a female exposure speaker and two test speakers, one female and one male), by morphing clear tokens of /s/ and /f/ by different proportions. Results from a pre-test categorization task showed that the male test speaker's ambiguous productions were perceived to be more /s/-like than those of the female exposure speaker, whereas overall perceptual responses for the two female speakers were similar. As a result, listeners generalized the adjusted representation of the fricatives from the female exposure speaker to the female test speaker, but not to the male test speaker, although all three speakers had perceptibly distinctive voices. In addition, when only a subset of ambiguous fricatives of the male test speaker were presented in the test stimuli such that the exposure and test stimuli were perceived to be equally /s/-like, learning of the exposure female talker generalized to the male test speaker. Results from this study indicate a role of bottom-up similarity (in terms of the extent of ambiguity, i.e., the degree of /s/-likeness) in guiding generalization in that generalization between two speakers was turned on and off by experimentally manipulating the sample perceptual space of the test speaker.

In another study, the generalization effect was tested on the perception of words (instead of a nonword-nonword continuum) produced by a novel talker (Witteman et al., 2013). Native-Dutch listeners who had limited prior experience with German-accented Dutch were briefly exposed to a German-accented speaker producing critical words with Dutch vowel /æy/ (pronounced as /ɔɪ/) in them. A post-exposure cross-modal priming task revealed that auditory primes with accented /æy/ productions facilitated recognition of identical visual targets, suggesting adaptation to the specific talker. However, when tested with a different speaker who had similar pronunciations of the vowel, prior exposure did not immediately facilitate word recognition, although it appeared to expedite the adaptation process.

In sum, research on cross-talker generalization is limited and has not revealed consistent

results with some showing adaptation restricted to specific speakers (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Jongman et al., 2003) and other showing adaptation effects readily transferred to other speakers (e.g., Kraljic & Samuel, 2006). The mechanism by which listeners generalize their prior experience with foreign-accented speakers to novel speakers remains poorly understood, although a few general suggestions have been made. Specifically, Kraljic and Samuel (2007) emphasized the patterning of talker information with phonetically-relevant acoustic cues. For instance, spectral cues that distinguish a /s/ from a /ʃ/ also reveals important information about talker (e.g., gender) and are more diagnostic of talker identity than temporal cues used for stops; thus, listeners tend to adapt in a talker-specific manner for fricatives but not stops. Reinisch and Holt (2014) provided evidence in support of perceptual similarity of talkers in constraining talker generalization. Bradlow and Bent (2008) suggested that systematic commonalities shared among a group of talkers evoke talker-independent adaptation, although they did not have confirmatory evidence to pinpoint the 'systematicity' at the acoustic-phonetic level (see also Sidaras et al., 2009). In Chapter 4, I first refine these general suggestions into testable working hypotheses. In particular, I make a distinction between the role of top-down influences (such as expectations for talker accent) and bottom-up input in constraining generalization. Then I present Experiments 4 and 5 which examine talker generalization of phonetic adaptation in two conditions: a) generalizing from a single foreign-accented talker to another talker with the same accent; and b) generalizing from a group of talkers who share an accent to a novel talker with the same accent.

Limits of Phonetic Adaptation: What Kind of Speech Input is Required?

Apparently, for non-native speakers, a major source of acoustic variability comes from the instability of L2 phonetic categories, of which productions are inevitably subject to

influences from native language (L1) phonology. Differences in phonetic inventories and in the specific realization of same phonemes across languages are known to challenge the phonetic learning of a foreign language and its production (e.g., Best & Tyler, 2007). Despite ample evidence of rapid phonetic adaptation to unfamiliar pronunciations, surprisingly little research has investigated how speaker-related factors, such as intelligibility, the degree of accentedness, and acoustic variation, may affect the adaptation process. These speaker-factors tend to vary substantially across foreign-accented speakers (e.g., Munro & Derwing, 1995) and will likely have direct influences on listeners' adaptation to the accent.

A Role of Speaker Intelligibility

In a foreign accent listening study, Bradlow and Bent (2008) compared native English listeners' speed of adaptation to Mandarin-accented speakers as a function of individual speakers' baseline intelligibility in English. Evidence indicates that while sentence transcription training improves recognition of accented sentences for all speakers, it takes a longer time for listeners to adapt to speakers of relatively low intelligibility, hypothetically due to weaker support for lexical-to-phonetic feedback in less intelligible speech. Indeed, past research shows that high-level linguistic information guides phonetic retuning of specific phonemes (e.g., Norris et al., 2003; Eisner et al., 2013). If an accumulation of retuned representations for several phonetic categories together improves sentence-level recognition for accented speech, a logical result of less intelligible speech is that more training sentences are required for adaptation, given fewer speech instances with clear lexical information in each sentence. However, other acoustic-phonetic level factors, such as larger within-talker variability in production, or more deviation from the acoustics of native tokens, may slow down adaptation processes and are more likely to occur in low intelligible speakers.

A Role of Within-Talker Variability

A number of studies that investigate native phonetic perception show that large within-talker variability, which increases the likelihood of acoustic overlap between categories, leads to slower responses in phoneme categorization and increased competition between phonetically-similar competitors in word recognition (Newman et al., 2001; Clayards, Tanenhaus, Aslin, & Jacobs, 2008;Toscano & McMurray, 2010). In addition, Hazan, Romeo, and Pettinato (2013) found that even after controlling for categorical overlap, within-category dispersion itself positively correlates with response speed in category identification. Few studies have explored the role of within-talker variation in foreign-accented speech and its influence on speech perception, which occurs above and beyond deviations from native acoustic distributions. Wade et al. (2007) reported poorer training effects in the recognition of foreign-accented spoken words, partially attributable to larger within-talker variability in non-native speech compared to native speech. However, in this study, because listeners were trained in a high-variability paradigm and were exposed to multiple speakers producing hundreds of words, it is hard to determine to what extent the adaptation was hindered by within- versus inter-talker variability independently, and whether listeners were puzzled by the overall stimulus variability across categories or by within-category variability of specific phonemes.

Studies of phonetic adaptation in foreign accents often do not report speaker intelligibility or within-talker variability of their productions, leaving open the question whether these factors have consequences on phonetic adaptation. One exception is Sumner (2011), which showed that some variability is better than none in eliciting phoneme-level retuning. Specifically, invariant tokens of French-accented /p/ (tokens with constant VOTs across phonetic contexts), which sounded like /b/ in English, did not generate typical recalibration of /b/-/p/ phonetic boundary ,

whereas exposure to an acoustically variable set of accented /p/ tokens did produce recalibration. Taken together with findings of Wade et al. (2007), these results suggest that adaptation to foreign-accented speech requires some, but not excessive, acoustic variability. In Chapter 5, I present Experiment 6 and 7 as case studies of talker-specific phonetic adaptation as an attempt to characterize the type of speech input that limits fast perceptual learning. Specifically, the role of speaker intelligibility and within-talker variability (of specific phonemes) is examined.

Characteristics of Mandarin-Accented English

As reviewed above, there is mounting evidence showing that listeners not only adapt to specific speakers by adjusting acoustic-phonetic mappings, they also sometimes generalize the remapping to novel talkers. The advance of theories lies in constructing a framework that captures the flexibility in the speech perceptual system and also specifies how the right degree of abstraction is achieved. Foreign-accented speech, with its noticeable deviation from native norms and its regularities across talkers who share an accent, provides an ideal test case to address questions about the mechanism of rapid phonetic adaptation and its generalization. In all of the experiments, I tested native-English listeners' adaptation to Mandarin-accented English, focusing on /d/ productions in word-final position.

In Mandarin, all stops are phonetically voiceless (no /b/, /d/, /g/ are found in Mandarin) and are distinguished by aspiration, instead of voicing. That is, the two tokens [t^h] and [t] are contrastive in Mandarin, but are allophones of the same category /t/ in English (Rochet & Fei, 1991). Moreover, Mandarin does not permit any stops in word-final position. Due to this L1 influence, voiced word-final stops (e.g., /d/ as in 'seed') are often devoiced (pronounced similar to [t^h]) in Mandarin-accented English and are perceptually confusable with voiceless tokens when judged by English listeners (e.g., Flege et al., 1992). Mandarin-accented English further

differs from native-accented English at the level of acoustic cues. In general, word-final voicing can be signaled by multiple acoustic cues: for instance, voiced stops tend to have longer preceding vowels, shorter closures and shorter bursts than unvoiced stops (e.g., Denes, 1955; Lisker, 1957; Hillenbrand et al., 1957). Native American-English speakers show salient and reliable vowel lengthening before voiced tokens across different phonetic contexts (Luce & Charles-Luce, 1985), but typically do not produce audible release of the final stop (Crystal & House, 1988). In contrast, Mandarin-accented speech demonstrates reduced or absent differences in vowel and closure durations, but usually keeps clear distinctions in burst durations (e.g., Bent, Bradlow, & Smith, 2008; Flege et al., 1992). Consistent with their differences in production of the voicing of stops, English and Mandarin listeners differ in their use of temporal cues to identify voicing in stop consonants (e.g., Crowther & Mann, 1992). The appropriate use of informative cues has been linked to enhanced intelligibility of foreign-accented speakers (Xie & Fowler, 2013).

Across all experiments, I combined acoustic analysis with behavioral responses in order to have a full understanding of adaptation to accents. Attending to acoustic detail might help to explain why in some situations listeners adapt and sometimes they do not; and why they generalize to some talkers but not others. In the studies reported in this thesis, all Mandarin-accented speakers were recruited from University of Connecticut. Speaker intelligibility was assessed by a pilot intelligibility study. Detailed information for the pilot study and demographic information of all speakers are presented in Appendix A. Following the perceptual learning paradigm, each experiment included an exposure phase and a test phase. Novel items and/or novel talkers were used at test to examine generalization of perceptual adaptation. Table 1

presents talker conditions across all experiments, which used cross-modal priming as the test task. Experiment 2 and 3 used different tasks and are not included in the table.

Table 1. Talker conditions across experiments.

Experiment	Condition	Exposure speaker(s)	Test speaker
Experiment 1-3	Talker-specific learning	Speaker 1	Speaker 1
Experiment 4	Generalization from a single talker	Speaker 1	Speaker 2
		Speaker 2	Speaker 1
Experiment 5A	Generalization from multiple talkers	Multi 1 (speaker 2,3,4,5,6)	Speaker 1
		Multi 2 (speaker 1,3,4,5,6)	Speaker 2
Experiment 5B[1]	Generalization from multiple talkers	Multi 1 (speaker 2,3,4,5,6)	Speaker 1
Experiment 6	Talker-specific learning	Speaker 2	Speaker 2
Experiment 7	Talker-specific learning	Speaker 3	Speaker 3

[1] Experiment 5B was conducted due to different generalization patterns in Experiment 5A. Details are presented in Chapter 4.

CHAPTER 3 THE SCOPE OF PHONETIC REORGANIZATION:

ADAPTATION RESHAPES INTERNAL STRUCURE OF PHONETIC CATEGORIES

Classic findings on categorical perception have revealed that while the acoustic signal is continuous, listeners efficiently map speech sounds onto phonetic categories that help to distinguish one word from another (Liberman et al., 1967). Though findings from categorical perception experiments suggest that listeners are not sensitive to variation within a phonetic category, findings from other paradigms suggest that members of the same phonetic category are not perceptually equivalent (e.g., Pisoni & Tash, 1974). As noted in Chapter 2, phonetic categories have a rich internal structure: both category membership and typicality of speech instances matter in speech perception (e.g., Andruski et al., 1994).

The literature on perceptual learning for speech shows that listeners use lexical information to disambiguate phonetically ambiguous speech sounds, and maintain this new mapping for later recognition of ambiguous sounds for a given talker. Evidence for this kind of perceptual reorganization has focused on phonetic boundary shifts. Here I present three experiments examining whether listeners adjust both category boundaries and internal category structure in rapid adaptation to foreign accents, and whether these phonetic adjustments, if any, help to alleviate lexical competition between phonetically-similar competitors.

Experiment 1 examined the effect of perceptual learning on spoken word recognition. Generalization of learning across the lexicon was examined by exposing participants to one set of words and testing them on a novel set. This study replicated the methods of Eisner et al. (2013) and extended this design to also ask whether perceptual adaptation results in changes in lexical competition. With a successful replication in Experiment 1, Experiments 2 and 3 were

designed to provide a precise examination of perceptual changes throughout the phonetic category that lead to improved word recognition. Experiment 2 examined changes in the phonetic category boundary using a category identification task. Experiment 3 examined influences of learning on the internal structure within the phonetic category by assessing listeners' goodness ratings of speech tokens as exemplars of each phonetic category (/d/ or /t/). Lastly, behavioral data were pooled across experiments and analyzed in combination with acoustic patterns of accented tokens in order to determine whether the re-weighting of acoustic cues contributed to rapid perceptual adaptation to the foreign accent.

Experiment 1

In Experiment 1, I investigated whether native listeners can rapidly adapt to Mandarin-accented word-final /d/ pronunciations. Two groups of native-English listeners were exposed to naturally-produced Mandarin-accented speech in an auditory lexical decision task during exposure. The experimental group heard a set of critical /d/-final words that were devoiced in the Mandarin-accented speech, but the control group heard only replacement words that did not contain any example of /d/. During test, all listeners completed a cross-modal priming task. The current design was modeled after Eisner et al. (2013) with one modification: we examined not only how auditory /d/-final words primed visual targets in an identity priming procedure (e.g., *seed –SEED*; visual targets are presented in capital letters), but also how they primed phonological competitors of the intended targets (e.g., *seed –SEAT*). I hypothesized that exposure to a novel accent would increase the match between accented input and lexical forms, resulting in larger identity priming effect for Mandarin-accented /d/-final words by the experimental group compared to the control group; in addition, following learning, intended targets would have

greater lexical activation than unintended competitors. Thus, the priming effect served as a measure of accent adaptation.

Methods

Participants. Forty-eight monolingual English speakers with no hearing or visual problems (according to self-report) were recruited from the University of Connecticut community. All participants were undergraduate students who were naïve to the Mandarin language and had no or minimal previous exposure to Mandarin-accented English. Participants were randomly assigned to one of the two exposure groups (experimental vs. control), with 24 participants in each condition. In this study, as in all subsequent experiments, participants received course credit or monetary reward for their participation, and gave informed consent according to the guidelines of the University of Connecticut Institutional Review Board.

Speech materials. One male native-Mandarin speaker with medium intelligibility (as determined by an intelligibility pilot study) was selected as the exposure and test speaker (Speaker 1; see Appendix A for demographic and intelligibility information). All words were produced naturally by this speaker. Recordings were made in a sound-proof room using a microphone onto a digital recorder, digitally sampled at 44.1 kHz and normalized for root mean square (RMS) amplitude to 70 dB SPL.

Exposure. For the experimental group, the exposure list consisted of 30 critical /d/-final words (e.g., *overload*), 60 filler words, and 90 nonwords. The list was identical for the control group except for the critical words. Instead of /d/-final words, there were 30 replacement words for the control group. The replacement words (e.g., *animal*) were matched to the critical -/d/ words in syllabic length and mean lemma frequency in CELEX (Baayen, Piepenbrock, & Gulikers, 1995). All words or nonwords were multisyllabic and contained three to four syllables.

In both conditions, auditory words were selected to meet the following criteria: 1) /d/ appeared only in word-final position, and only in critical words; 2) no other alveolar stops, no other voiced stops or dental fricatives, and no post-alveolar affricates occurred; 3) no voiceless stops (/p/ or /k/) occurred in word-final position. The same criteria were used in the selection of test stimuli.

Test. The test list was identical for both exposure groups. There were 60 monosyllabic /d/-final words (taken from /d/-/t/ minimal pairs such as "*seed-seat*") and 180 monosyllabic filler words. Mean lemma frequencies in CELEX of the /d/- and /t/-final items were 83 (SD = 186) and 87 (SD = 126) per million, respectively, $t(59) = .159$, $p = .88$.

Procedure. Each participant completed an auditory lexical decision task during exposure, which was immediately followed by a cross-modal priming task. A between-subjects design was used such that, during the exposure phase, the experimental group and the control group heard items from the experimental list and the control list, respectively. Items were presented in a random order. For the auditory lexical decision task, participants were instructed to decide whether each auditory stimulus was a real English word and to press a corresponding button as quickly and accurately as possible.

The test phase was identical for both groups. Participants were told that they would continue to hear auditory words (primes) but immediately after that they would see visual letter strings (targets) presented on the screen. The task was to decide with a yes/no button press whether the visual stimuli were real English words or not. On critical trials, 60 words from /d/- and /t/-final minimal pairs appeared as visual targets, in four different prime –target pairing types: /d/-final words as visual targets preceded by an identity prime (e.g., *seed –SEED*) or an unrelated prime (e.g., *fair –SEED*); /t/-final visual targets preceded by a minimal pair contrast (e.g., *seed –SEAT*) or an unrelated prime (e.g., *fair –SEAT*). Successful encoding of the accented

/d/ variant should be manifested as a greater magnitude of priming for identity priming ([*seed –SEED*] – [*fair –SEED*]) in participants who heard /d/-final words than for those who listened to the accented talker but heard no /d/-final tokens.

Words in each set of minimal pair items were rotated over four counterbalanced lists and within each list, they appeared in only one of the four prime –target pairing type. Each counterbalanced list had equal proportions in the four pairing types. Non-critical trials were identical across counterbalanced lists: 30 filler words were paired with an identical prime (e.g., *foam –FOAM*) or an unrelated prime (e.g., *male –HORN*), and another 90 auditory filler words were paired with visual nonwords (e.g., *ring –WELF*). Thus, among the 180 trials in each list, half the targets were nonwords. The test lists were pseudo-randomly ordered such that no more than four words or nonwords appeared in a row, and the critical trials were evenly spaced. For each list, there were two test orders in which trials were in reverse order. Example stimuli across counterbalances lists are presented in Table B1 (Appendix B).

Stimuli were presented using Eprime 2.0.8 running on a desktop computer. Audio stimuli were delivered via Sennheiser HD280 headphones at a comfortable listening level constant across participants; visual targets were shown in white Helvetica font in lower case on a black background in the center of the computer screen. During exposure, ten practice trials were given to the participants before the actual task to familiarize them with the task procedure. Practice items were similar to filler words and did not appear in the exposure stimuli. Exposure auditory items were presented with an inter-onset interval of 3000 ms. During test, ten practice trials of the cross-modal priming task were given to participants, followed by the actual test. The inter-trial interval was 1400 ms, timed from the button press response to the onset of the next auditory prime. Visual targets were presented immediately at the offset of the auditory prime and stayed

on the screen for 2 s unless terminated by a response. Reaction times (RT) were measured from visual target onset. During both phases, participants were told to respond as fast as possible without sacrificing accuracy. Responses were made via keyboard with two buttons labeled ‘yes’ and ‘no’. Assignment of the ‘yes’ button to the right or left hand was counterbalanced across participants.

Results

Exposure. Response accuracy is presented in Table C1. Of interest, critical /d/-final words were largely judged to be real words by the experimental group (*M* = .81, *SD* = .09).

Test. Table C2 shows mean error rates and reaction times (RT) in the test phase. Analysis of error rates did not reveal any group differences and were omitted from discussion here. RTs for correct responses were analyzed. Items were discarded from the statistical analysis if the rate of correct identification across all participants was less than 41% accurate (three standard deviations (SD = 16%) below the mean (89%) across all items). By this criterion, three words (*plod, moot, spate*) were discarded in this experiment and in all experiments presented in Chapters 3 -5. In addition, a preliminary inspection revealed that extreme outliers in the RTs caused a violation of the normality assumption of the RT data. Responses above or below 2 SDs from the mean of each prime type in each group were excluded from the RT analysis (4.5% of correct trials). Fig.1 shows the RT priming magnitude (unrelated minus related) as a function of *exposure group* and *target type*.

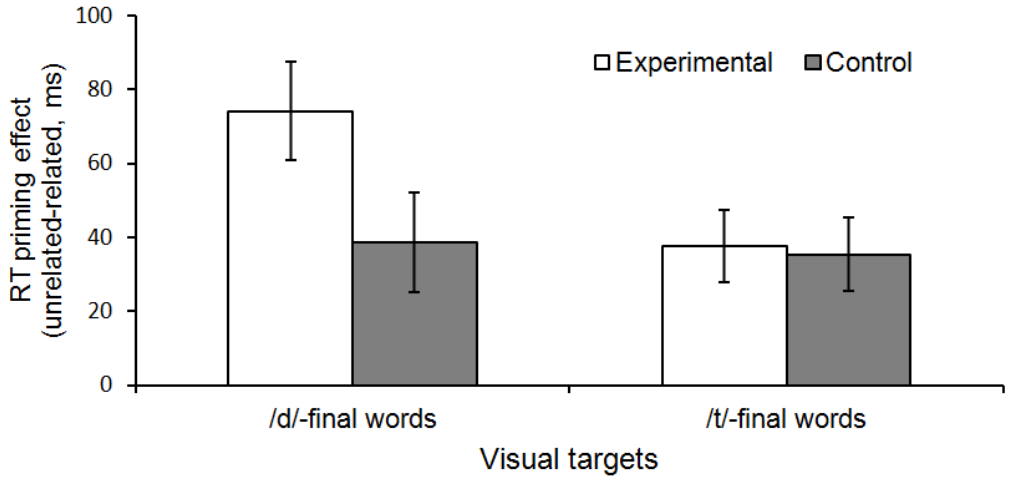

Fig.1. Experiment 1 test results: Priming of /d/-final words (RT in *fair-SEED* trials minus RT in *seed-SEED* trials) and /t/-final words (RT in *fair-SEAT* trials minus RT in *seed-SEAT* trials) for participants exposed to critical words (Experimental group) or replacement words (Control group). Error bars represent standard errors of the mean.

I report the RT results analyzed in a linear mixed-effects model. The model included *exposure group* (experimental vs. control), *target type* (/d/-final vs. /t/-final words), *prime type* (related vs. unrelated primes) and their interactions as fixed effects. Random effects included by-subject intercepts and by-item intercepts and slopes for priming type, which had the maximal random effect structure justified by the data[2] (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013). All the independent variables were contrast coded (contrast coding is the default coding in ANOVA; the interpretation of coefficients are comparable to main effects in ANOVA) as follows: *exposure group*: experimental = 1, control = -1; *target type*: /d/-final targets = 1, /t/-final targets = -1; *prime type*: related = 1, unrelated = -1. The coefficients reflect the distance of each level of the variable from the overall mean of the variable. As expected, related primes elicited faster responses than unrelated primes (β = -23.23, SE = 2.48, $p < .0001$).

[2] The random effects in all models reported in the dissertation were determined by a stepwise variable selection procedure. The reported model always contained the maximal random effect structures justified by the data. I used the lme4 package in R (Bates, Maechler, Bolker & Walker, 2014) to conduct the analysis.

Meanwhile, /d/-final targets elicited slower responses than /t/-final targets (β = 13.57, SE = 4.88, $p < .01$). Of interest, there was a significant three-way *exposure group* × *target type* × *prime type* interaction (β = -4.11, SE = 2.11, $p < .05$). No other effects were significant at the .05 level.

Given the interaction, we further analyzed the data by fitting mixed-effects models for /d/-final targets and /t/-final targets separately, with *exposure group*, *prime type*, and their interaction as fixed effects. For /d/-final targets, there was a significant priming effect (β = -27.44, SE = 3.27, $p < .0001$); crucially, the priming effect was significantly larger in the experimental group than the control group, as revealed by the interaction effect (β = -7.58, SE = 3.24, $p < .05$). Thus, relative to control participants, participants in the experimental group who had been exposed to /d/-final words showed larger identity priming (e.g., "*seed-SEED*") during test. Notably, critical test words were not heard during the exposure phase. Therefore, gains in identity priming likely reflect increased compatibility between accented tokens and phonetic representations, which generalized across the lexicon and facilitated subsequent word recognition. This result replicated findings of Eisner et al. (2013); they found an increase of identity priming as native-English listeners adapted to final-devoiced /d/ in Dutch-accented English. We added two novel types of critical trials (e.g., "*seed-SEAT*" and "*fair-SEAT* ") to Eisner et al.'s design to examine to what extent accented /d/ tokens activate the representation of /t/. For /t/-final targets, there was a priming effect (β = -19.31, SE = 3.84, $p < .0001$), suggesting that responses made to /t/-final words (e.g., "*SEAT*") were faster following an auditory /d/-final word (e.g., "*seed*") than following a phonologically unrelated word (e.g., "*fair*"). Importantly, there was no *exposure group*-by-*prime type* interaction (β = .46, SE = 2.73, $p = .87$), suggesting that there was no group difference in terms of the absolute priming magnitude for /t/-final targets.

Thus, exposure to critical /d/-final words increased priming for /d/-final targets in the experimental group, without decreasing priming for the voiceless competitor, /t/-final targets.

We further asked whether within each exposure group (experimental vs. control), auditory /d/-final words elicited larger priming for the intended item (e.g., "*seed-SEED*") than for the phonological competitor (e.g., "*seed-SEAT*"). Starting with the control group, there was a significant priming effect (β = -18.79, SE = 2.87, $p < .0001$) but no *prime type*-by-*target type* interaction (β = -0.19, SE = 2.87, $p = .95$). The absence of an interaction effect indicated that these listeners, who had not heard any examples of /d/ during exposure, activated the lexical representation of both the intended words (e.g., "*seed*") and their phonological competitors (e.g., "*seat*") almost equally. Thus, as predicted, Mandarin-accented /d/-final words were overall somewhat ambiguous for untrained native-English listeners. In contrast, in the experimental group, the main priming effect (β = -26.98, SE = 3.65, $p < .0001$) was modulated by a *prime type*-by-*target type* interaction (β = -8.28, SE = 3.65, $p < .05$), reflecting larger priming for the intended item (e.g., "*seed-SEED*") than for the phonological competitor (e.g., "*seed-SEAT*").

Discussion

In summary, two important results emerged from this experiment. On the one hand, in the experimental group, perceptual learning did enhance lexical activation of the intended target ("*seed-SEED*") such that intended lexical items received higher levels of activation than their competitors that differed by a voicing feature. On the other hand, the lack of group effect for /t/-final targets indicated that the experimental group (as well as the control group) exhibited significant priming from -/d/ primes on -/t/ targets (e.g., "*seed—SEAT"*). Thus, among experimental participants who had exposure to critical -/d/ words before test, even though /d/-final words were more strongly activated (e.g., "*seed—SEED"*), this increased activation was not

at the expense of activation of /t/-final words (e.g., "*seed—SEAT"*). The influence of perceptual learning on the competing lexical target (e.g., "*seed—SEAT*") was not assessed in Eisner et al. (2013) or other studies that showed successful adaptation to natural *foreign* accents (e.g., Witteman et al., 2013). I now compare the current results to previous studies that investigated perceptual learning of atypical pronunciations in one's *native* accent.

McQueen et al., (2006) showed a complete elimination of priming effect on phonological competitors (e.g., "*doos-DOOF*", both are words in Dutch) after listeners adapted to the ambiguous fricatives (midway between /s/ and /f/) embedded in a native Dutch accent. Similar results were obtained by Sjerps & McQueen (2010), who investigated the perception of a single non-native sound embedded in native speech. In this study, Dutch listeners adapted to a noncanonical /f/ or /s/ sound (actually replaced by the English /θ/ sound, as in "*bath*"). Importantly, the /θ/ sound elicited an identity priming effect of the same magnitude as that elicited by an unambiguous native sound, whereas no significant priming on phonological competitors were found. The results were taken as evidence of thorough learning of non-native sounds. However, our data indicated that the Mandarin-accented -/d/s did not fully function like native phonemes even after critical exposure.

The discrepancy between the current data and previous studies might arise for a number of reasons. First of all, in both McQueen et al. (2006) and Sjerps and McQueen (2010), the manipulated sound was the only unfamiliar sound that needed to be adapted to; the rest of stimuli were normal, clear native speech. In the current study, natural phonetic variation that deviates from the native norm was pervasive in the stimuli, in the sense that many other segments (e.g., vowels and other consonants) also bore traces of the non-native accent. Although perceptual learning has been shown to be largely automatic (Witteman, Bardhan, & Weber, 2014; but see

Zhang & Samuel, 2013), the requirement of simultaneous adjustment to multiple phonetic categories may change the time course of complete learning. Second, the differences in results might be due to the specific phoneme class being tested. Notably, fricatives (Kraljic & Samuel, 2005) elicit larger adaptation effects (as shown by a larger shift in the location of phonetic boundaries) than stop consonants do (Kraljic & Samuel, 2006). Lastly, the natural variation in -/d/ tokens (i.e., variation within the intended category) used in our study might also have made the task more demanding. Future work is needed to tease apart these possibilities. It is also important to test whether constraining the range of acoustic variation among exposure and test items (cf. Sumner, 2011) would help listeners to achieve complete learning of a foreign accent faster, or if complete learning is achievable at all during brief exposure. Such tests will help us to establish the limits of perceptual learning. In Experiments 2 and 3, I sought to provide a more precise indication of pre-lexical changes, by examining changes in the location of phonetic boundary between categories and in the internal structure within the categories.

Experiment 2

Previous studies of perceptual learning have measured the learning result in terms of shifts in phonetic category boundaries (e.g., Norris et al., 2003; Kraljic & Samuel, 2005). At test, listeners were generally asked to identify tokens that varied along an acoustic continuum. The test sounds were artificially created by mixing two clear sounds (e.g., /s/ and /f/) in different proportions. In natural speech, category membership is often determined not by a single acoustic dimension, but by the combination of multiple acoustic cues. Each acoustic dimension has its own distributional characteristics and is differentially informative about phonetic segment identity. These cues and their informativeness in foreign-accented speech can be quite distinct from those in the native speech (Flege et al., 1992); on the other hand, due to influences of L1

phonology, they usually vary systematically across talkers from the same language community (e.g., Bent et al., 2008). Hence, in order to capture in full scope any potential acoustic-phonetic level adaptation to rich variation in the accented speech, I used naturally-produced /d/-final and /t/-final words as test stimuli in Experiment 2, instead of mixing pairs of tokens in predefined proportions to create a continuum. Following an exposure phase that was identical to that in Experiment 1, I assessed potential changes in the phonetic category boundary using a two-alternative, forced-choice (2AFC) category identification task during test. A phonetic boundary shift would be indicated by an increase in /d/ responses for /d/-final words.

Methods

Participants. Forty-eight monolingual English speakers with no hearing or visual problems were recruited from the University of Connecticut community. Participants were naïve to the Mandarin language and had no or minimal previous exposure to Mandarin-accented English. Participants were randomly assigned to one of the two exposure groups (experimental vs. control) with 24 participants in each condition.

Speech materials. The exposure stimuli were identical to those used in Experiment 1. The test list included 60 monosyllabic minimal pairs ending in /d/ or /t/ (e.g., *seed –seat*; identical to the /d/-final words that appeared in Experiment 1 during test as auditory primes). The test stimuli were organized into two blocks such that for each participant, members of the same minimal pair did not appear in the same block. For example, if *seed* appeared in block 1, *seat* appeared in block 2. The order of blocks was counterbalanced across participants. Each block consisted of 30 /d/-final words and 30 /t/-final words; items were presented in random order within each block.

Procedure. Stimuli were presented using the same equipment as in Experiment 1. During both phases, participants were told to respond as fast as possible without sacrificing accuracy. The exposure phase was identical to that used in Experiment 1. During the test phase, test items were presented via headphones with an inter-trial interval of 2000 ms. Listeners were asked to identify the final consonant of each item as either /d/ or /t/ by pressing an appropriately labeled button. No feedback was provided.

Results

The categorization results showed that there was large variability (percent /d/ responses ranged from 0 to 100 for both /d/-final and /t/-final words) across items in terms of their ambiguity, as expected for naturally-produced non-native accented speech. A mixed-effects logit model was used to analyze the category identification data (Fig.2). Mixed logit models predict the probability of a particular response (here, a /d/ response; Agresti, 2002; Jaeger, 2008). In analyzing the current data, main effects of *exposure group* (experimental vs. control) and *word type* (/d/-final vs. /t/-final) as well as their interaction were included in the model. By-item intercepts and by-subject intercepts and slopes for word type were included as random effects, which had the maximal random effects structure justified by the data. The independent variables were contrast coded as follows: *exposure group*: experimental = 1, control = -1; *word type*: /d/-final = 1, /t/-final = -1. For the dependent measures, /d/ responses were coded as 1 and /t/ responses were coded as 0. Positive log coefficients indicate a log odds ratio greater than 0 (corresponding odds ratio is greater than 1), which means that the level coded as 1 has greater probabilities of /d/ responses than the level coded as -1. Overall, /d/-final words (64%) elicited significantly more /d/ responses than /t/-final words (32%) across the two groups (log coefficient $\beta = 1.00$, SE = .16, $p < .0001$). Crucially, there was a significant group effect ($\beta = .22$, SE = .09,

$p < .05$): The experimental group reported significantly more /d/ responses (51%) than the control group (45%) overall. There was no *exposure group*-by-*word type* interaction ($\beta = .08$, SE = .08, $p = .28$). The category identification results taken as a whole indicate that the experimental group tended to interpret more words (both /d/-final and /t/-final) as ending in /d/ than the control group, suggesting a boundary shift towards the /t/-end along a /d/-/t/ continuum.

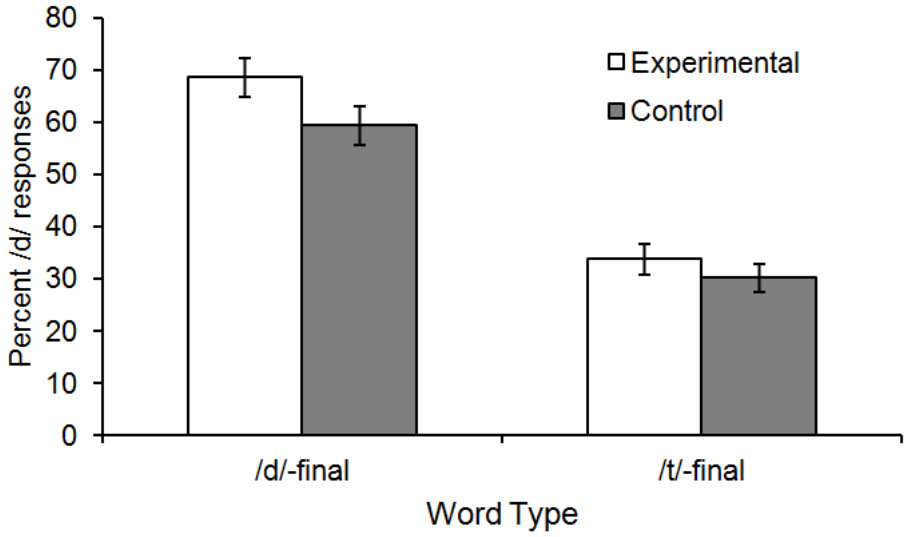

Fig.2. Mean percent /d/ responses for the 2AFC category identification task in Experiment 2 as a function of exposure group and word type. Error bars indicate standard errors of the mean.

Discussion

Experiment 2 replicated previous findings on rapid perceptual learning of ambiguous sounds (Norris et al., 2003; Kraljic & Samuel, 2005; Reinsich & Holt, 2014): there was an increase in identification of noncanoncial sounds as members of the trained category. We did not observe any group-by-word type interaction, suggesting that the learning is likely driven by the absorption of ambiguous tokens near the boundary into the /d/ category, rather than by enhanced discrimination between /d/ and /t/. This pattern did not emerge in previous studies due to the fact that when listeners were exposed to the ambiguous sounds, they were also hearing clear tokens

from the contrastive category (see also Sumner, 2011). In that type of acoustic environment, broadening of one category does not compromise the identification accuracy of words containing tokens from another category since there is no acoustic overlap between categories. In natural foreign accents, however, between-category acoustic overlap is often observed for both vowels (e.g., Wade et al., 2007; Sidaras et al., 2009) and consonants (e.g., Warner et al., 2004). If all perceptual learning does is to incorporate ambiguous sounds into one category without instigating changes with respect to perceived goodness of the accented tokens in general, then a great amount of difficulty may persist with native listeners when they encounter tokens of confusable categories. In Experiment 3, I address this issue directly by asking whether foreign-accented tokens become better instances of the intended category following perceptual learning.

Experiment 3

The objective of Experiment 3 was to investigate potential learning consequences on the internal structure of phonetic categories. Perceived goodness of accented tokens was assessed by a goodness rating task, which tapped into listeners' sensitivity to phonetic detail in a more graded way than categorical membership. Even for unambiguous tokens, the perceived goodness can be adjusted due to contextual influences such as speaking rate (Volaitis & Miller, 1992; Allen & Miller, 2001). I asked if perceptual learning of a non-native accent would have similar influences on the phonetic structure by changing the way in which fine-grained phonetic variation in foreign-accented tokens are perceived by native listeners. If so, I expected to see group differences with respect to goodness ratings of accented tokens as exemplars of the intended phonetic categories.

Methods

Participants. Forty-eight monolingual English speakers with no hearing or visual problems were recruited from the University of Connecticut community. Participants were naïve to the Mandarin language and had no or minimal previous exposure to Mandarin-accented English. Twenty-four participants appeared in the experimental and the control conditions, respectively.

Speech materials. The exposure and test stimuli were identical to those used in Experiment 2.

Procedure. The exposure phase was identical to that used in Experiments 1 and 2. Following the exposure phase, ten practice trials were given to each participant before the start of the test phase. The practice helped to ensure that listeners were rating goodness of the final consonant of each item, not the degree of accentedness of the whole word. During the practice, ten monosyllabic words ending in /m/ were randomly presented[3] and listeners were asked to focus on the final sound of each auditory item and rated its goodness as /m/. They were asked to rate each item on a scale from 1 to 7, with 7 being a very good exemplar of the category and 1 being a very poor exemplar. We reasoned that despite of the strong accent in all words, if a participant gave high ratings for words ending in clear /m/ but low rating for words which were perceived to end in /n/, then he/she understood the task. An experimenter was present during this practice phase to make sure that participants followed the procedure.

During the test phase, the test stimuli were divided into two sets and were administered in two blocks. The words from a minimal pair did not occur within the same block. In order to examine how exposure to /d/-final words influenced perceived goodness of /d/ tokens as

[3] The words were recorded by the test speaker and administered to a separate group of native-English listeners in a word transcription task. Results from the transcription task indicated that the coda consonant ranged from being a very clear /m/ (good examples of /m/) to a clear /n/ (bad examples of /m/); all words were also strongly-accented as a whole (i.e., contained strongly-accented vowels or word-initial consonants).

members of the /d/ category, participants were asked to rate the final consonant of each item in terms of how good it was as an exemplar of a /d/ by pressing an appropriately labeled button. In order to examine potential influences from the critical exposure on the representation of phonetic category on the other end of the voicing continuum, we also asked listeners to rate each item for goodness as /t/. Goodness as /d/ served as the primary dependent measure, we consider results from goodness as /t/ a source to provide complementary information about the perceptual changes along the entire voicing continuum. Participants rated goodness as /d/ in one block and goodness as /t/ in another block. The allocation of test sets and the order of blocks were counterbalanced across participants, such that half the participants rated goodness as /d/ first. Each block consisted of 30 /d/-final words and 30 /t/-final words; within each block, items were presented in a random order. Participants were asked to rate each item on a scale from 1 to 7, with 7 being a very good example of the category and 1 being a very poor example. Auditory items were presented with an inter-trial interval of 2000 ms. No feedback was provided.

Results

To accommodate individual variability and potential rating bias (Schütze & Sprouse, 2011) and to understand the relative rating of each item, participants' raw ratings were transformed into standardized z-scores that were used in subsequent analyses. Fig.3 presents the mean standardized ratings for each task as a function of exposure condition. Note that for each rating task (goodness as /d/ and goodness as /t/, separately), this within-subjects standardization procedure makes each participant's mean rating across all test items zero; the mean rating for /t/-final words is necessarily the additive inverse of that for /d/-final words. Thus, I only present the mean rating for /d/-final words in the goodness-as-/d/ task and mean rating for /t/-final words in the goodness-as-/t/ task. However, in the statistical analysis as reported below, both /d/-final

words and /t/-final words were included for each task because all the reported effects were taken as random at the item level.

Goodness as /d/. A linear mixed-effects model was fitted with *exposure group*, *word type* and *group-by-word type* interaction as fixed effects. By-item intercepts were included as random effects. A contrast coding scheme was used for independent variables as in Experiment 2. Goodness ratings as /d/ revealed a main effect of *word type* ($\beta = .42$, SE = .04, $p < .0001$), indicating that /d/-final words received higher ratings than /t/-final words. There was no effect of *exposure group* ($p = .99$). The learning effect took the form of an interaction between *exposure group* and *word type* ($\beta = .05$, SE = .01, $p < .001$): relative to the control group, the experimental group rated /d/-final words as better examples of /d/ and rated /t/-final words as poorer examples of /d/.

Goodness as /t/. A similar linear mixed-effects model was fitted to analyze goodness ratings as /t/. /t/-final words received higher ratings than /d/-final words ($\beta = -.48$, SE = .05, $p < .0001$). There was no group effect ($p = .99$). Again, the learning effect was revealed in the *group*-by-*word type* interaction ($\beta = -.04$, SE = .01, $p = .005$): relative to the control group, the experimental group rated /t/-final words as better examples of /t/ and rated /d/-final words as poorer examples of /t/.

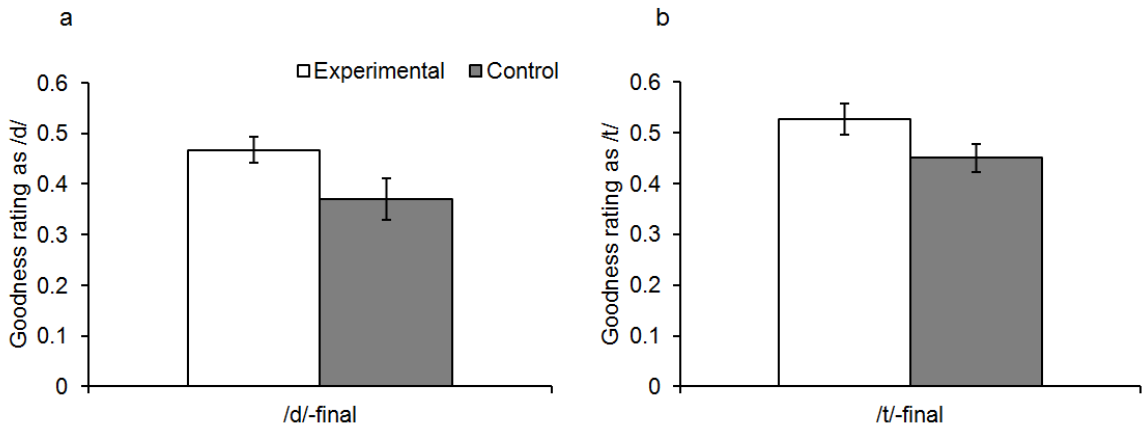

Fig.3. Mean goodness ratings (z-transformed) of each word type (/d/-final words and /t/-final words) as exemplars of (a) /d/ category and (b) /t/category as a function of each exposure group in Experiment 3. Error bars indicate standard errors of the mean.

Discussion

In both goodness rating tasks, the experimental group (relative to the control group) assigned higher ratings to words in accordance with the intended word type. Crucially, although the experimental group was exposed to /d/-final words only, the exposure affected their judgment of speech tokens as exemplars of the /t/ category: /t/-final words were perceived to be better /t/s among trained participants than untrained control participants. Note that the categorization results from the control group in Experiment 2 showed that our test stimuli varied over a wide range in their ambiguity. Some items fell unambiguously into the unintended category; some fell into the ambiguous niche; others were clear, albeit non-native, tokens of the intended category. If whatever perceptual changes following exposure were limited to the boundary region, we would not be likely to observe a global improvement in the perceived goodness of speech tokens as members of the intended category (/d/ or /t/). In fact, in Experiment 2, exposure to /d/ productions led to more /d/ responses to both /d/-final and /t/-final words in categorization. If those ambiguous tokens near the boundary that drove the change in the boundary location also

drove the group difference in the perceived goodness, we would observe /t/-final words would be rated as 'better' as members of the /d/ category. In contrast, in Experiment 3, the experimental group gave higher goodness as /t/ to /t/-final words than the control group. Consider this seemingly inconsistent pattern, it is possible that the /d/-tokens that drove up the report of /d/ responses in the identification task were those closer to the boundary region, whereas the /d/-tokens that led to higher goodness ratings were those closer to prototypical variants of the phonetic category. The fact that an overall improvement of perceived goodness was found in aggregate across the /d/ category and that the learning effect extended to untrained phonetic category (/t/) suggests that listeners were not merely incorporating /d/ tokens into the intended segmental category (e.g., Norris et al., 2003). But rather, in concert with a phonetic boundary shift, listeners also adjusted the internal structure within each phonetic category for the /d/-/t/ contrast.

Phonetic Adjustment: a Re-Weighting of Acoustic Cues

Experiment 1 showed that native-English listeners adapted to Mandarin-accented /d/-final words. Perceptual learning effects generalized across the lexicon: novel accented words elicited larger lexical activation among listeners who were previously exposed to the specific phoneme than among control listeners. Experiment 2 and 3 together revealed that perceptual adaptation to the Mandarin-accented speaker had consequences for the global phonetic structure of the alveolar stop contrast in word-final position: it did not only lead to a broadened /d/ category among the experimental group, compared to control participants (Experiment 2), but also affected the perceived goodness of these tokens as an exemplar of relevant phonetic categories (/d/ and /t/; Experiment 3). Previous studies of talker-specific perceptual learning have emphasized that listeners alter the location of between-category boundary as a result of

experience with the productions of a particular speaker (e.g., Norris et al., 2003; Reinisch & Holt, 2014). Experiment 2 replicated this finding. Note that a shift in the phonetic category boundary itself does not entail changes in perceived goodness of within-category tokens (Allen & Miller, 2001). Listeners adjust the internal structure of a phonetic category only when they detect systematic changes in how a segment is uttered, for example, in the case of changing speaking rate (Miller & Volaitis, 1989), or place of articulation (Volaitis & Miller, 1992). In Experiment 3, I provided the first evidence that listeners do so during rapid perceptual learning of a specific talker's accent.

One interesting but somewhat unexpected result was that exposure to /d/ tokens affected the phonetic representation of the /t/ category. Presumably, representations of internal structure of the two categories could be independent from each other. Then why is there a "carryover" effect? To answer the question, we need to find out what perceptual dimensions are involved in such adjustments. Most theories of speech recognition assume that multiple cues are involved in speech categorization: either they are stored in fine-grained detail in memory and compared to incoming speech during recognition (e.g., Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2006); or alternatively, they are integrated in a multi-dimensional space into an abstract form to inform phoneme categorization (e.g., McClelland & Elman, 1986; Norris et al., 2000). Although it is implied in perceptual learning studies that listeners are sensitive to talker-specific distribution of acoustic cues (e.g., Kraljic & Samuel, 2006, 2007), few studies have investigated the specific acoustic-phonetic properties associated with learning, other than implying that listeners learned general information such as "this speaker produces odd /d/ tokens" (Kraljic & Samuel, 2006) or "lowered vowels" (Maye et al., 2008). Thus, relatively little is known about the exact informational source of the sound-to-category remapping process (see Reinisch, Wozny,

Mitterer, & Holt, 2014 for an examination of specific acoustic cues in visually-cued phonetic recalibration). Outside the domain of talker-related perceptual learning, studies of speech categorization suggest that listeners are generally sensitive to the statistical values of critical acoustic properties in the speech input (e.g., Clayards et al., 2008) and weight cues differentially as a function of their informativeness in distinguishing phonetic categories (Toscano & McMurray, 2010). More importantly, training with category-level feedback can shift listeners' attention to more informative acoustic cues over less informative ones as they learn non-native phonetic contrasts (Francis, Baldwin, & Nusbaum, 2000; Francis & Nusbaum, 2002). Here I consider a similar mechanism that may underlie listeners' reorganization of phonetic structure of /d/ and /t/ in adapting to the Mandarin-accented speaker.

As described in Chapter 2, Mandarin-accented English differs from native accents at the level of acoustic cues. These characteristics were well-manifested in the current test stimuli. Fig.4 presents the distributional pattern of the acoustic cues: Durational differences in vowel and closure durations were uninformative in cueing voicing, whereas the difference in the burst release was striking. Not only was the final stop released for every token (both /d/ and /t/), but also bursts contained durational information that could be used to reliably differentiate voiceless from voiced tokens, with only very small overlap between the two categories. I thus hypothesized that adaptation to the accent, and in particular, the adjustment of internal structures of /d/ and /t/ categories, is achieved via an adjustment in the weighting of various acoustic cues for the accent. To test this hypothesis, I assessed behavioral responses in the experimental and control groups of Experiments 2 and 3 as a function of the acoustic properties of the speech materials.

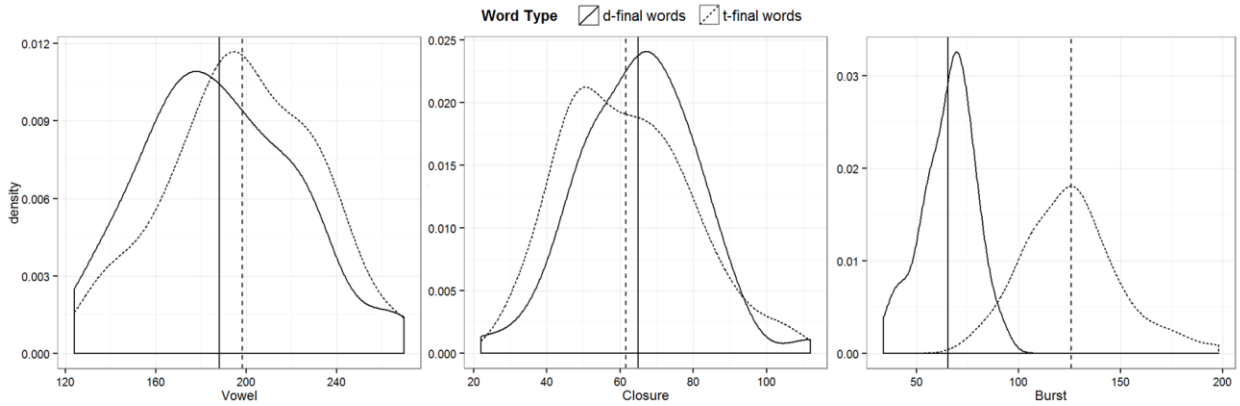

Fig.4. Density plots of acoustic measures (preceding vowel duration, closure duration and the length of burst and aspiration, respectively) across all 60 minimal pairs (/d/-final vs. /t/-final words) used in the test phase of Experiment 2 and 3. Vertical lines show the mean value for each word type.

Mixed-effects regression models were fitted to predict the categorization and goodness judgment responses by including *exposure group* and the three temporal measures (duration of *vowel*, *closure* and *burst*) as well as their interactions (between exposure group and each acoustic measure) as predictors. Subjects and items were considered random effects. Predictors were standardized before they were entered into the regression model. The predictive power of the acoustic cues reveals how informative they are (i.e., the perceptual weighting) in determining phonetic membership and category typicality; the interaction with exposure group reveals changes, if any, in the relative weighting as a result of exposure to critical words. I present the regression results for the category identification and goodness ratings separately (Table 2).

Category Identification (Experiment 2)

There was a main effect of *exposure group*, with the experimental group reporting more /d/s throughout the acoustic continuum ($p < .05$). The main effects of all three measures were significant: vowel duration, $p < .05$; closure, $p < .001$; burst, $p < .0001$. Although none of the interaction terms were statistically significant, there was a trend ($p = .10$) for the experimental group to rely more on the burst in making categorization decisions than the control group.

Table 2. Results across Experiments 2 and 3: estimated probability of /d/ responses (Experiment 2) and goodness of /d/ and /t/ tokens (Experiment 3) as a function of temporal acoustic cues. Numbers in parentheses represent standard errors.

	Experiment 2		Experiment 3			
	Probability of /d/ responses		Goodness as /d/		Goodness as /t/	
Predictor	Log Coefficient	p value	Coefficient	p value	Coefficient	p value
Intercept	- 0.08 (0.16)	.63	-0.002 (0.04)	.96	-0.002 (0.05)	.97
Group	0.21 (0.09)	.02*	-5E05 (0.01)	.99	-1E04 (0.01)	.99
Vowel	0.29 (0.14)	.03*	0.09 (0.04)	<.05*	-0.05 (0.05)	.35
Closure	-0.54 (0.14)	<.001**	-0.20 (0.04)	<.0001**	0.18 (0.05)	<.001**
Burst	-0.93 (0.14)	<.0001**	-0.42 (0.04)	<.0001**	0.43 (0.05)	<.0001**
G × Vowel	-0.05 (0.03)	.15	-0.01 (0.02)	.50	-1E04 (0.01)	.99
G × Closure	-0.04 (0.04)	.25	-0.01 (0.02)	.52	-0.008 (0.01)	.59
G × Burst	-0.06 (0.04)	.10	-0.03 (0.01)	.05*	0.03 (0.01)	.01*

Goodness Rating (Experiment 3)

Goodness as /d/. Both closure and burst information predicted goodness rating as /d/ (*p*s < .0001); the main effect of vowel duration was also significant ($p < .05$). Interestingly, there was a significant group-by-burst interaction ($p = .05$); the signs of coefficients suggest that the experimental group relied on this acoustic property more heavily than the control group. No other effects were significant.

Goodness as /t/. Both closure ($p < .001$) and burst information ($p < .0001$) predicted goodness rating as /t/, but the main effect of vowel duration was not significant ($p = .35$). As in the preceding analysis, there was again a significant group-by-burst interaction ($p = .01$),

indicating a heavier weighting of this acoustic dimension by the experimental group relative to the control group. No other effects were significant.

Together, evidence suggested a heavier weighting of burst information by the experiment group across all three tasks, although goodness ratings were more sensitive than categorization responses in detecting changes in the cue-weighting functions. The results from the control group provided a gauge of initial cue use without exposure to the critical words: listeners were generally sensitive to multiple cues, even the ones that were not typically used for native contrasts (i.e., burst length in cueing word-final voicing, Hillenbrand et al., 1957). For the experimental group, the variation of acoustic dimensions of critical words present during exposure further guided their attention to the most informative cue when tested with novel stimuli. The group-by-burst interaction on perceived goodness clearly suggests that the internal structure of the phonetic categories were reorganized as a result of a re-weighting of acoustic parameters. As noted above, burst length potentially provides reliable information to distinguish voiced tokens from voiceless ones in the productions of this particular Mandarin-accented speaker. Increased attention towards this acoustic dimension could explain the unexpected adjustment in the internal structure of the /t/ category. Evidence of a shift in cue-weighting strategy refines and expands our understanding of the cognitive mechanism underlying the rapid sound-to-category remapping process: Listeners are not only capable of tracking acoustic distributions of a single acoustic dimension in adapting to unfamiliar pronunciations (e.g., voice onset time, see Sumner, 2011) but also readily re-weight their reliance on different acoustic cues in making phonetic decisions. This result parallels findings of cue-weighting changes in second language acquisition, although learning of non-native phonetic contrasts requires more training and occurs over longer time scale (see Francis et al., 2000). Taken together, these results

illustrate a highly flexible perceptual system that is context-sensitive: the informational integration across multiple acoustic dimensions is tailored to the particular speaker (accent) or language. It should be noted that since we used natural speech tokens, we did not have rigorous control over the acoustic cues, nor was the current study designed to warrant a precise calculation of cue weights. Future studies should use an orthogonal design to assess how categorization and goodness judgments change across one cue while holding the other cues constant (e.g., Holt & Lotto, 2006) or examine the relative cue-weighting when cues signal conflicting information (Francis et al., 2000). Investigations in this direction would further elucidate how adaptation to talker-related characteristics arises by integrating over multiple cues (even cross-modally, see Reinisch et al., 2014).

CHAPTER 4 THE MECHANISM OF CROSS-TALKER GENERALIZATION: ACOUSTIC SIMILARITY SUPPORTS GENERALIZATION TO NOVEL TALKERS

Two domains of research have separately investigated whether listeners generalize the adaptation to a specific speaker to other speakers. The literature on phonetic recalibration of specific segments, with a focus on adaptation to artificially-created ambiguous sounds (e.g., Norris et al., 2003; Kraljic & Samuel, 2005, 2006, 2007), implicates a discrepancy for specific phoneme classes. Namely, spectrally-shifted fricatives tend to elicit talker-specific adaptation, whereas temporally-cued stops are found to elicit talker-independent adjustments. Researchers have noted that the spectral cues that are used to distinguish fricatives tend to vary more substantially across talkers (Newman et al., 2001), whereas the temporal cues that distinguish voicing stops from voiceless ones are less predictable by talker information (Allen et al., 2003). It is unclear what the discrepancy reflects: "bottom-up constraints" that are specific to the speech signal; or, "top-down expectations" (guided by long-term experience) to encode acoustic-phonetic properties in a more talker-specific manner if talker-identity characteristics tend to be present in the altered segment itself (e.g., fricatives, vowels). For instance, do listeners generalize for stop consonants because the specific acoustic attributes (e.g., VOT) that cue phoneme identity were indeed highly similar across tested speakers (as in Kraljic & Samuel, 2006, 2007), or because their long-term experience with the native language enables them to *infer* that two speakers would likely have similar productions of stops and thus they readily applied the same sound-to-category mapping? Previous studies have not been able to dissociate these two possibilities (e.g., Kraljic & Samuel, 2007).

Reinisch and Holt (2014) provided support for the bottom-up similarity account: phonetic recalibration for fricatives was restricted to a specific speaker when a novel speaker was not perceptually similar in segmental productions, but generalized when the two speakers were similar. However, without corroborating evidence from stop consonants that shows a similar dissociation, we still do not know whether there is a general tendency for listeners to be more conservative for some kinds of phonetic adjustments. In addition, as noted in Chapter 2, bottom-up similarity was measured by listeners' responses in a categorization task (whether tokens were /s/-like or /f/-like) in Reinisch and Holt (2014). It remains an open question exactly at which type of sub-lexical level listeners were generalizing: phoneme category (e.g., 'ambiguous sounds are /f/s') or specific acoustic cues (e.g., 'spectral centroid within this range denotes /f/'). Of note, Witteman et al. (2013) selected two speakers who both substituted the Dutch vowel /æy/ with German vowel /ɔɪ/, yet no immediate generalization was observed between the speakers. It is possible that more fine-grained acoustic differences between the two speakers hindered the generalization.

This distinction also speaks to the findings from studies of intelligibility in non-native accents. This literature shows that on the one hand, listeners are unable or reluctant to generalize across talkers based on experience with a single speaker; on the other hand, exposure to a group of talkers promotes generalization (e.g., Bradlow & Bent, 2008; Sidaras et al., 2009), although not consistently (e.g., Wade et al., 2007; Clarke, 2000). A lack of phonemic and sub-phonemic measures in these studies makes it hard to accurately interpret the contributing sources of cross-generalization when it occurs. Problematically, the use of different paradigms in the exploration of conditions for talker generalization makes it difficult to compare the generalization following

multi-talker exposure in intelligibility studies to the generalization between single talkers in studies tapping into phonetic recalibration of specific categories.

In this chapter, I present a set of experiments examining cross-talker generalization. I have two specific goals. The *first* goal is to differentiate between a *top-down expectation* hypothesis vs. a *bottom-up similarity* hypothesis to account for the discrepant findings regarding generalization of phonetic retuning. To this end, I asked whether listeners would generalize in a talker-independent manner for stop consonants, which have been previously found to elicit talker-general adaptation in native accents (Kraljic & Samuel, 2005, 2007), when listeners perceive a natural foreign accent. In practice, listeners may ascribe perceived acoustic-phonetic variation to idiosyncratic or talker-general sources and demonstrate different generalization patterns accordingly. For instance, phonetic recalibration of category boundary is only observed when acoustic variation is attributed to speaker idiosyncrasies, but not when it is part of a context-conditioned dialectal feature (Kraljic, Brennan, & Samuel, 2008) or a consequence of incidental factors (e.g., a pen in the mouth, Kraljic, Samuel, & Brennan, 2008). If top-down expectations play a role in constraining talker generalization, they may be applied differently in presence of an unfamiliar foreign accent. The *second* goal is to validate the hypothesis that multiple-talker exposure benefits talker generalization by allowing talker-independent retuning of *specific phonetic categories*. Note that natural variation both within- and across- talkers serves a double role: as a cue to phoneme identity and as a cue to talker accent information. I have a particular interest in the interaction between bottom-up acoustic-phonetic structures and listeners' perception of a shared accent (among multiple talkers) in constraining generalization.

I used the paradigm of Experiment 1 to examine cross-talker generalization of accent learning in two exposure conditions: single-talker exposure and multiple-talker exposure.

Experiment 4 (*single* talker condition) investigated whether exposure to a talker's non-native accent generalizes to the phoneme of interest (word-final /d/) for a different talker with the same accent. Experiment 5 (*multiple* talker condition) further tested whether listeners show generalization to a new talker following exposure to a group of talkers that share the same accent. In each experiment, I combined acoustic analysis with listeners' behavioral performance as well as their subjective reports of talker similarity in order to pinpoint the mechanism that subserves talker generalization of phonetic recalibration.

Experiment 4

Previous research on native accents shows that phonetic retuning operates in a talker-specific manner for fricatives but in a talker-independent manner for stop consonants. It is unclear whether the asymmetry in results for stops and fricatives was due to top-down expectations of the patterning of speaker specificity for stops versus fricatives, or bottom-up similarity/dissimilarity present in the specific speech signal, or even inherent processing differences for temporal versus spectral cues. In Experiment 4, I examined a different test case: adaptation to stop consonants in unfamiliar foreign accents. Due to first language (L1) influences, productions of L2 speakers contain noticeable acoustic deviations from native norms of the L2. Three alternative working hypotheses are developed. First, if there are processing differences for spectral vs. temporal cues such that listeners always encode temporal cues in a talker-independent manner regardless of who is talking, we would replicate talker-independent adaptation for stop consonants in a natural foreign accent. Second, if listeners use top-down expectations to constrain generalization, then we would not find generalization across talkers unless listeners' explicit judgment of the situation warrants it. For example, even though listeners have a tendency to generalize atypically pronounced stop consonants across native-accented

talkers, they may refrain from generalizing when they do not have a good estimate of whether their prior experience applies (e.g., when noticing speakers are of unfamiliar non-native accents), unless they believe the speech input comes from the same person or same accent. Third, if talker-generality of phonetic retuning for stops in previous studies is the consequence of acoustic similarity across talkers, then we would find evidence for talker generalization only when the exposure talker and test talker are acoustically similar in their productions of the critical *segment* (Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2014). Of note here, Reinisch and Holt (2014) reported generalization between talkers even when the talkers had perceptually different *voices* and were identified as different voices by listeners.

The design of Experiment 4 followed Experiment 1, consisting of an exposure phase and a test phase. The only difference was that speech materials for the test phase were now produced by a novel Mandarin speaker. I asked if listeners' prior experience with the exposure talker's pronunciations of /d/-final words (e.g., *overload*) affects subsequent recognition of novel /d/-final words (e.g., *seed*) and their voicing minimal pairs (e.g., *seat*) when produced by the test talker, by comparing the priming effects in the experimental group versus the control group. Improved spoken word recognition for the test talker in the experimental group would suggest cross-talker generalization of adjusted phonetic representation of /d/ category. Upon the completion of behavioral tasks, listeners were asked to identify whether they noticed a talker change between the two phases; if they did, they were further asked to rate the accent similarity of the two talkers. I integrated the results from these questions into the analysis of behavioral data.

Methods

Participants. Fifty-two undergraduate or graduate students from University of Connecticut participated in this experiment. One early English-Spanish bilingual was excluded. Three additional participants were excluded for poor performance during the exposure phase (response accuracy below or at chance level). Forty-eight participants were included in the analyses, with equal numbers of participants in the experimental and the control group (n = 24 each). All participants were monolingual English speakers with no hearing or visual problems. According to self-reports at the end of the experiment, all participants had no or minimal prior experience with Mandarin-accented English or the Mandarin language.

Speech materials. Two male native-Mandarin speakers (Speaker 1 and Speaker 2) with equivalent intelligibility (as determined by a pilot intelligibility study) recorded speech stimuli for this experiment (See Appendix A for demographic information). We found evidence of talker-specific adaptation in Experiment 1 for Speaker 1. Here, with each exposure group, for half the participants, Speaker 1 served as the exposure talker and Speaker 2 was the test talker (Speaker 1 → Speaker 2); another half of the participants heard Speaker 2 as the exposure talker, and Speaker 1 as the test talker (Speaker 2 → Speaker 1). Thus, in each of two groups (experimental and control), twelve participants heard speaker 1 as the exposure talker. This design helped to control for any asymmetry in talker generalization originating from talker peculiarities. Materials were identical to those used in Experiment 1 and all words were recorded and digitally processed in the same procedure as in Experiment 1.

Procedure. The experimental procedure of the exposure and test phase was identical to that in Experiment 1. After participating in the behavioral tasks, listeners were asked to indicate whether they noticed a talker change between phases. If their answer was “Yes” , they were further asked to rate the voice similarity and accent similarity of the speakers on a scale from 1 to

7, with 7 being identical and 1 being very different. Participants were specifically instructed to rate the accent similarity in terms of the type of accent (language community), rather than the strength of accentedness.

Results

Exposure. Response accuracy (collapsed across talkers) is presented in Table C1. Again, critical /d/ words were largely judged to be real words by the experimental group (M = .83, SD = .07). Accuracy for critical /d/ words did not differ between the two speakers, $t(22) = 1.637$, p =.11. Thus, for both speakers, we expected that their speech tokens during exposure should provide enough lexical information to elicit an adjustment in the phonetic representation of /d/.

Test. Responses (4.9% of correct trials) above or below 2 SDs from the mean of each prime type in each exposure group were excluded from the RT analysis. Table C3 shows mean error rates and reaction times (RT) in the test phase. Of interest was the magnitude of priming (unrelated minus related) as a function of exposure group and target type (Fig.5).

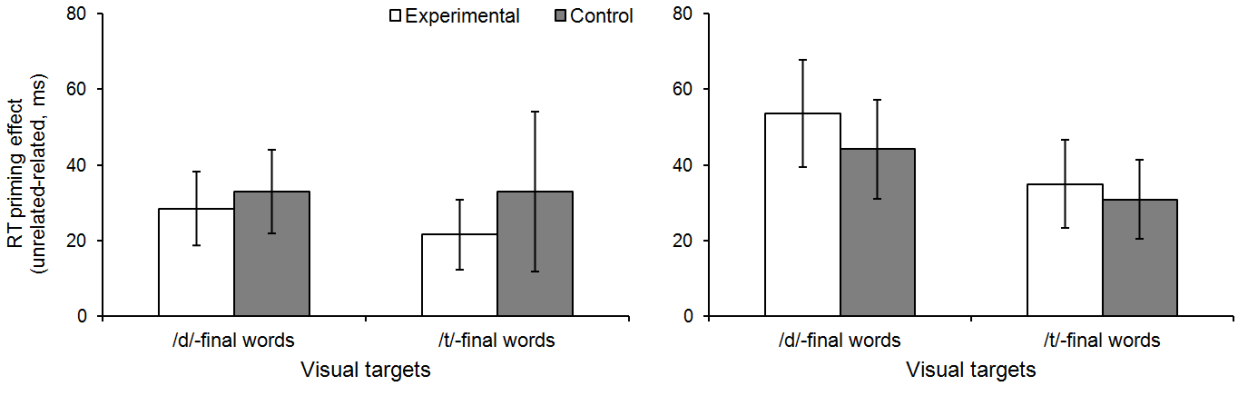

Fig.5. Experiment 4 test results: Speaker 1 → Speaker 2 condition (left panel) and Speaker 2 → Speaker 1 condition (right panel). Priming of /d/-final words (RT in *fair-SEED* trials minus RT in *seed-SEED* trials) and /t/-final words (RT in *fair-SEAT* trials minus RT in *seed-SEAT* trials) for participants exposed to critical words (Experimental group) or replacement words (Control group). Error bars represent standard errors of the mean.

A mixed-effects model was fitted with RTs as the dependent measure. Both fixed effects and random effects were the same as in Experiment 1. There was a significant priming effect (β = -16.77, SE = 2.20, $p < .0001$). Meanwhile, /d/-final targets elicited slower responses than /t/-final targets (β = 15.05, SE = 4.77, $p < .01$). There was no interaction between *target type* and *prime type* (β = -1.08, SE = 2.20, $p = .62$), suggesting that Mandarin-accented /d/-final word (e.g., *seed*) activated the intended word equivalently to the close phonological competitor (/t/-final targets, e.g., seat). This was expected for Mandarin-accented /d/ productions which are often perceived as /t/ tokens by native-English listeners. Crucially, unlike in Experiment 1, there was no three-way *exposure group* × *target type* × *prime type* interaction (β = -1.17, SE = 2.17, $p = .59$). Thus, there was no influence of exposure group on the priming magnitude for either /d/-final or /t/-final targets, suggesting that exposure to one speaker's production of critical /d/ words did not improve recognition of /d/-final words produced by a different speaker; that is, cross-talker generalization was not observed. This suggested that different generalization patterns for fricatives versus stops in past research were unlikely due to processing differences for spectral vs. temporal cues.

Nevertheless, it is possible that generalization between talkers was not symmetrical such that exposure with one speaker transferred to another speaker but not vice versa. Such an asymmetry might obscure any evidence of an overall cross-talker generalization. Another mixed-effects model was fitted, with *talker condition* (Speaker 1 → Speaker 2 vs. Speaker 2 → Speaker 1) *exposure group, target type* and *prime type* as well as their interactions as fixed effects. No main effect of *talker condition* was found (β = .12, SE = 8.05, $p = .99$), and *talker condition* did not interact with *prime type* in any way ($ps > .10$). The results suggested the priming pattern was not affected by the specific exposure/test talker.

I further statistically assessed whether participants' response patterns differed as a function of their reports of talker and/or accent similarity. 12 out of 24 participants in the experimental group identified the exposure talker and test talker as the same person; rated voice similarity of the two speakers by participants who perceived two voices was 3.82 (*SD* = 1.89; range = 1-6) on a scale of 1-7 (7 = same voice; 1 = very different voices). 12 out of 24 participants in the control group identified the exposure talker and test talker as the same person; rated voice similarity by participants who perceived two voices was 4.58 (*SD* = .67; range = 3-5) on a scale of 1-7. *Voice judgment* (same speaker vs. different speakers) as a binomial factor (contrast coded as follows: same speaker = 1, different speakers = -1) was included into the mixed-effects model. The model included *exposure group*, *target type*, *prime type*, *voice judgment* and their interactions as fixed effects. Results revealed no main effect of *voice judgment* (β = -8.95, SE = 9.05, *p* = .33). Of particular interest, *voice judgment* did not interact significantly with other factors either (*p*s > .10). Thus, even when listeners believed that the test and exposure talkers were the same person, no generalization was observed.

A similar analysis was conducted on the priming patterns with respect to individual participants' accent judgment. In both groups (experimental and control), 13 out of 24 participants identified the exposure talker and test talker as having the same accent. The average ratings of accent similarity by participants who perceived different accents was numerically higher in the control group (Experimental: *M* = 3.73, *SD* = 1.07, range = 2-5; Control: *M* = 4.75, *SD* = 1.22, range = 3-7) on a scale of 1-7 (7 = same accent; 1 = very different accents). A mixed-effects model was fitted, including *accent judgment* as a binomial factor (same accent vs. different accents; contrast coded as follows: same accent = 1, different accents = -1). Again, There was no main effect of *accent judgment* (β = -5.05, SE = 8.98, *p* = .58). There was no

significant interaction between *accent judgment* and other factors (*p*s > .05), suggesting that the perception of accent similarity between the speakers did not affect the generalization pattern.

Given that even listeners who identified the talkers as having the same voice and listeners who grouped the talkers as speaking with the same accent did not show any sign of talker generalization, I suspect that a lack of generalization might be due to a lack of bottom-up support from acoustic cues. I conducted acoustic analyses on the critical-/d/ words produced by the two speakers, focusing on three temporal cues that were found to be diagnostic of voicing in English stops: preceding vowel duration, closure duration and the length of burst and aspiration of the stop (Table 3). I compared the exposure words (3-4 syllables) and test words (monosyllabic) separately, considering that word length substantially changes the duration of temporal acoustic cues (Klatt, 1976; Lehiste, 1972). Independent samples t-tests showed that the two speakers had significantly different production patterns for the exposure words: speaker 2 had significantly longer vowels ($t(58) = 4.428$, $p < .001$), longer closures ($t(58) = 6.450$, $p < .001$) and longer bursts ($t(58) = 6.263$, $p < .001$) than speaker 1. Speaker 2 also produced the test words with significantly longer bursts than speaker 1, $t(118) = 3.505$, $p < .001$, whereas the acoustic realizations of vowels and closures showed the same trend but not significantly (*p*s > .10). Note that Speakers 1 and 2 were selected as test pairs because they were matched on overall intelligibility, and their productions of /d/-final words were of equivalent ambiguity (somewhat /t/-like) in a pre-test (see Appendix A for detail). For the critical test /d/-final words, 65% (*SD* = 13) of speaker 1's productions and 70% (*SD* = 6) of speaker 2 were identified as ending in /d/ in a 2AFC task (/d/ or /t/) in the intelligibility pilot study. To speculate, a trading relation among vowel duration, closure duration and burst duration could explain why the acoustic patterns differed between the speakers but the overall intelligibility was similar. Specifically, consider /d/

tokens of speaker 2: while the relatively long duration of vowels might lead to the perception of a voiced token, long closure and long burst could be cues to a voiceless token. Since speech perception is the result of integration across multiple acoustic dimensions, the overall interpretation of these cues might have made speaker 2 as intelligible (or unintelligible) as speaker 1. Overall, the data suggests that listeners might be reluctant to generalize when there is acoustic misalignment between speakers. I now situate current finding in the context of past research and discuss our interpretation of the data in relation to alternative hypotheses of talker generalization in detail.

Table 3. Mean (standard deviations) acoustic measures (in msec) by speaker, averaged across critical /d/-final items.

Phase		Vowel	Closure	Burst
Exposure	Speaker 1	148 (34)	37 (18)	79 (12)
	Speaker 2	194 (46)	68 (19)	116 (30)
Test	Speaker 1	188 (34)	65 (16)	66 (13)
	Speaker 2	198 (38)	64 (16)	77 (23)

Discussion

In the current experiment, no difference was observed during the test phase between the experimental group and the control group. Despite prior exposure to a talker who produced /t/-like /d/ words, the experimental group did not recognize critical test /d/-final words any better than the control group when the words were produced by a different talker with the same foreign accent. Moreover, for both groups, an auditory /d/-final word led to lexical competition between minimal pairs of /d/-/t/ words, without favoring either one. The results were in direct contrast with our previous finding of talker-specific learning. In Experiment 1, exposure to a Mandarin-accented speaker elicited phonetic retuning of /d/ category such that listeners were more likely to

recognize an auditory /d/ token as /d/ than /t/. I took the current finding as evidence for an absence of generalization across talkers and ruled out the first hypothesis that listeners maintain a single sound-to-category mapping for stops across all talkers.

The results were also in contrast with the findings of Kraljic and Samuel (2006, 2007), who showed that listeners generalized perceptual learning of stop consonant categories (e.g., /d/ - /t/) between two native talkers (one male and one female). Of note, critical items for the two speakers in their study were acoustically close on a number of measures including closure and burst duration. In a computational model, Mirman, McClelland and Holt (2006) tested the hypothesis that talker generalization patterns were directly linked with inter-talker acoustic similarity of phonemically-distinctive features. Consistent with human data, when the acoustic similarity of critical cues was high across speakers (e.g., "burst" and "voiced" features for stops), simulation data showed cross-talker generalization; when the acoustic similarity was low along critical feature dimensions (for fricatives), there was no talker generalization. However, the difference between stops and fricatives were still confounded with the amount of inter-talker acoustic variability in this study. Solid support for the hypothesis requires additional evidence of double dissociation: talker generalization for fricatives when talkers are sufficiently similar and no generalization for stop consonants when exposure talker and test talker do not have bottom-up similarity for phonetically-distinctive acoustic cues. Reinisch and Holt (2014) showed support for the fricative portion of this dissociation: listeners generalized their experience of a female speaker's ambiguous fricative productions (/f/ or /s/) to a male speaker only when the productions of the two speakers were perceived to be similar (as measured in their acoustic ambiguity between /f/ and /s/). The current results provided evidence for the second half of that dissociation: although listeners may generally tend to adapt to unfamiliar pronunciations of stop

consonants in a talker-independent way, they did not generalize experience from a Mandarin-accented speaker's stop consonants to a novel Mandarin-accented speaker when the acoustic patterns misaligned between talkers.

Interestingly, in Reinisch and Holt (2014), generalization was observed despite the fact that listeners judged the two speakers to have different accents and there was clearly no confusion between voices. Similarly, Experiment 4 indicated that the lack of generalization was not affected by listeners' explicit perception of talker voices or accents. The acoustic analysis of the Mandarin-accented speakers' /d/-final words was consistent the notion of *acoustic similarity* as a constraint of cross-talker generalization: despite the finding that half the participants did not detect a talker change, the acoustic patterns of the critical words were very different between the two speakers and listeners did not generalize across speakers. With evidence across previous studies and the current work, I favor the third hypothesis (bottom-up similarity) over the second one (top-down expectations) as an account for talker generalization.

The lack of cross-talker generalization in the current experiment aligned with findings from other paradigms on foreign-accented speech, which consistently reported that training on words spoken by one foreign-accented speaker did not improve intelligibility of other speakers (e.g., Bradlow & Bent, 2008; Jongman et al., 2003). Our analysis suggests that as foreign-accented speakers transfer their L1 phonology to the target L2, the realization of specific phonemes could be inconsistent across speakers; such inconsistence might have constrained listeners from generalizing across talkers in previous studies. Put simply, while speakers of Mandarin may share the same general accent in English, the way this accent is manifested can vary significantly across segments. Moreover, when productions of the specific phonemes are acoustically dissimilar, listeners are unwilling to apply their belief that the talker is the same at

exposure and test, or that the talker shares the same accent to generalize what they have learned about /d/ to this new talker. Similarly, we can imagine that in other situations where listeners may have more accent knowledge (for instance, given sentence-level stimuli, Bradlow & Bent, 2008), such knowledge itself is not sufficient to override a bottom-up mismatch.

Based on our data, it is fair to say that the perceptual system operates conservatively to the extent that listeners do not generalize what they learn from one speaker to another, at least when the bottom-up acoustic cues do not support generalization. This leads to a question why multiple-talker exposure has been shown to lead to cross-talker generalization in some situations (e.g., Sidaras et al., 2009). Experiment 5A aims to provide some answers to this question.

Experiment 5A

While Experiment 4 showed a lack of generalization of phonetic learning from one talker to a new talker, evidence exists that listeners generalize from *multiple* talkers to one or more novel talkers (Bradlow & Bent, 2008; Sidaras et al., 2009). This evidence is taken to show that listeners can *extract* systematic information across multiple talkers to overcome talker-specific variation and make general adjustments transferrable to new members with the same accent. The presence of *multiple* talkers may reveal more information about a shared accent than when there is only a single exposure talker. I refer this as the "extraction" hypothesis. While it seems like a very plausible account, it is worth considering why listeners sometimes do not generalize from a group of talkers to new talkers. Highly comparable to the design of Sidaras et al. (2009), Wade et al. (2007) also trained native-English listeners with a group of Spanish-accented speakers over several days in word transcription tasks. Yet in this study, improvement of recognition was restricted to trained speakers. Acoustic analysis revealed high acoustic variability of the vowel inventories both within- and across-talkers. The researchers pointed out that this high variability

in non-native tokens was an obstacle to accent adaptation. This raises the question whether naturally-produced foreign accents can really provide listeners "systematic variability" that allows them to extract acoustic-phonetic structure at an accent (talker-independent) level.

In Experiment 5A, I set out to provide a more rigorous test of the *extraction* hypothesis. The same exposure-test paradigm as in Experiment 1 was employed. The only change was that words in each exposure condition were spoken by five different Mandarin-accented talkers instead of a single talker. A novel Mandarin-accented talker served as the test talker. Notably, there are a few differences between the design of previous research and the current study. First, instead of a word transcription task, I used the perceptual learning paradigm to track changes in the representation of a single phonetic category (/d/ in word-final position). At test, the cross-modal priming task allowed us to compare the relative activation of target words and phonological competitors. As noted in the introduction, in transcription tasks, listeners merely had to choose a word from the mental lexicon that provide the best match to the speech signal. For instance, if listeners heard [trIt] ("trit", a nonword) while the intended word was "treat" ([trit]), their lexical knowledge could help them to correct the non-native pronunciation and they would report the intended word. In the cross-modal priming task, however, an initial misperception would result in an increase of reaction time. In addition, no other phonemes confusable with the critical phoneme were presented in the speech stimuli throughout the experiment. In this way, if any generalization effects were found, we could be sure that improved word recognition is due to enhanced representation of the specific segment. Second, I collected subjective reports from participants in order to gauge whether the accents of talkers were indeed perceptually similar to the listeners. Both participants' explicit ratings and acoustic patterns of the critical words were taken into account when discussing results on generalization patterns.

Third, similar to Experiment 4, I created two pairings of exposure-test talkers to make sure that if we observe generalization from a group of talkers to a novel talker, it is not due to factors incidental to the test talker.

To make it a testable, I elaborate the *extraction* hypothesis into two scenarios. If cross-talker generalization reflects active abstraction across talkers guided by top-down expectations, then listeners must be aware at some level of the shared accent among talkers in the multiple-talker exposure conditions. In reality, top-down expectations could come from visual information of talker identity, or attention to other details that are not intrinsic to the specific segments (e.g., atypical stress patterns), among many others. In this case, we would see supporting evidence from participants' reports on accent similarity across talkers. It is an open question whether listeners have such awareness of accent type. Relevant to this question, Skoruppa & Peperkamp (2011) tested French listeners' adaptation to an artificially-created novel dialect of the native language by altering the vowel pronunciations (see also Maye et al., 2008). Their results indicated that listeners had explicit knowledge of the dialectal context where the utilization of shifted phonetic categories was appropriate. Other studies have shown that listeners are sensitive to difference in speakers' dialects and track them separately (Trude & Brown-Schmidt, 2012). Likewise, it is possible that as listeners are exposed to an unfamiliar foreign accent, they not only make online adjustments for specific segments, they also build up a representation of what the accent sounds like. The latter type of learning would provide listeners a basis to infer whether talkers are similar and help to constrain generalization when new talkers are encountered. A second possibility is that talker generalization is driven by bottom-up similarity (of the segment) among talkers, specifically by retuning listeners' attention to particular aspects of the segmental productions (for instance, certain regions in the perceptual space or specific acoustic dimensions)

that are stable across talkers. In this case, listeners' explicit awareness of a similar accent is not necessary. However, it is crucial that talkers show commonalities along acoustic cues that are distinctive for specific phonemes. If this is the case, we would see supporting evidence from the acoustic analysis of critical words.

Meanwhile, an alternative possibility should be noted. It is that previous findings of talker generalization in intelligibility studies might be accounted for by a general relaxation in the mapping from nonstandard speech signals to lexical representations without the mediation of an altered phonetic representation. Listeners are shown to be more tolerant of acoustic mismatches when speech tokens deviate from canonical forms and accept phonologically similar words as speech targets (Brouwer et al., 2012; McQueen & Huettig, 2012). A general relaxation may also account for the finding of Baese-Berk et al. (2013) showing that training with a group of foreign-accented speakers with various accents improved word recognition for an untrained accent (see discussion of Baese-Berk et al., 2013 for other explanations). I deem this "general relaxation" hypothesis unlikely given convincing data from past research and results from Experiment 1 that listeners engage in phonetic retuning to accommodate talker-specific unfamiliar pronunciations. However, if the "general relaxation" hypothesis is true, listeners should show increased activation for both target words (/d/-final words) and their phonological competitors (/t/-final words) upon hearing the critical /d/-final words.

Methods

Participants. Fifty-five monolingual English speakers with no hearing or visual problems were recruited from the University of Connecticut community. One early English-Italian bilingual was excluded. Six participants were excluded for poor performance in the exposure phase (response accuracy below or at chance level) or misunderstanding the test task.

Forty-eight participants were included in the following analyses, with equal numbers of participants in the experimental and the control condition.

Speech materials. In addition to the two native-Mandarin test speakers, four male Mandarin speakers (speakers 3-6) were selected from a previously recorded pool. All speakers were late L2 learners of English. A pilot intelligibility study suggested variability across speakers, with the two test talkers in Experiment 4 in the medium range (Appendix A). These two speakers again alternately served as test talkers in Experiment 5A. Two talker conditions were constructed: half participants were exposed to speakers 2, 3, 4, 5, and 6 and were tested with speaker 1 (Multi1 → Speaker 1); the other half of participants were exposed to speakers 1, 3, 4, 5, and 6 and were tested with speaker 2 (Multi 2 → Speaker 2). The exposure list was identical to that used in Experiment 1; the only change was that words in each exposure group were spoken by five different talkers with the number of items evenly divided between talkers. A pilot intelligibility study ensured that the overall ambiguity of critical /d/-final words during exposure was equated to that in Experiment 1. Recording and digital processing of all speech stimuli were completed following the same procedure as that in Experiment 1.

Procedure. The procedure was identical to that of Experiment 1. Following the exposure and test phase, listeners were asked to a) report the number of speakers in each phase; b) categorically indicate whether the accents of speakers (between exposure and test phase) were the same or not; c) rate the accent similarity between exposure talkers and test talkers on a scale of 1-7; d) guess accent type of the talkers if possible.

Results

Exposure. Data were collapsed across exposure talkers and are presented in Table C1. Response accuracy indicated that critical /d/-final words were largely judged to be real words by

the exposure group (M = .78, SD = .08). Accuracy for critical /d/ words did not differ between Multi 1 and Multi 2 speaker groups, $t(22) = 1.410$, $p = .17$.

Test. Table C3 shows mean error rates and RTs of correct responses in the test phase. Responses (5.3% of correct trials) above or below 2 SDs from the mean of each prime type in each exposure group were excluded from the RT analysis. A mixed-effects model was fitted on RTs of correct trials as in Experiment 1. There was a significant main effect of *prime type* (β = -16.55, SE = 2.17, $p < .0001$) and a main effect of *target type* (β = 10.70, SE = 4.40, $p < .05$). There was a significant interaction between *target type* and *prime type* (β = -5.55, SE = 2.16, $p < .05$), driven by a larger priming size for /d/-final words (41ms) than /t/-final words (22ms). Thus, accented /d/-final words were a better match to the intended lexical candidates (/d/-final targets, e.g., *seed*) than to phonologically similar competitors (/t/-final targets, e.g., *seat*). Note that in Experiment 4, the same test talkers were used and there was no difference between priming effects for /d/- vs. /t/- final targets. The larger priming sizes for /d/-final words here suggested some learning of the accent with respect to acoustic-phonetic variation of the /d/ category, instead of a general looser criterion in mapping sounds to words.

However, although the priming size for /d/-final targets was numerically larger in the experimental group than in the control group, the three-way *exposure group* × *target type* × *prime type* interaction was not significant (β = -1.53, SE = 2.05, $p = .46$). That is, priming patterns as a function of target type did not differ by exposure group (experimental vs. control). It was not expected the control participants to show any learning for /d/ tokens since they were not exposed to the critical sounds before test. I suspect that this surprising result was driven by different response patterns to the two test talkers. For instance, a lack of generalization to one of

the test talkers might have obscured an interaction. Or alternatively, if the result was meaningful, we wanted to know whether the learning effects held for both test talkers.

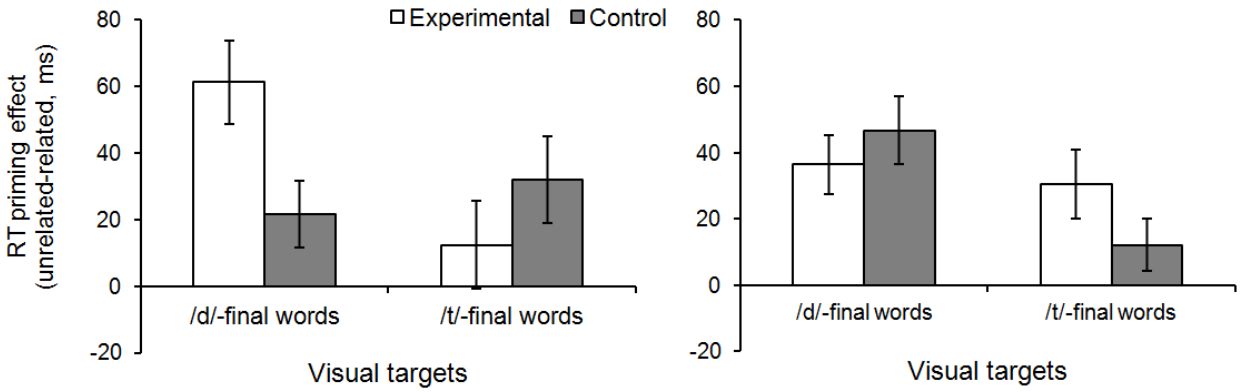

Fig.6. Experiment 5A test results: Multi 1 → Speaker 1 condition (left panel) and Multi 2 → Speaker 2 condition (right panel). Priming of /d/-final words (RT in *fair-SEED* trials minus RT in *seed-SEED* trials) and /t/-final words (RT in *fair-SEAT* trials minus RT in *seed-SEAT* trials) for participants exposed to critical words (Experimental group) or replacement words (Control group). Error bars represent standard errors of the mean.

Again, a mixed-effects model was fitted with *talker condition* (Multi 1 → Speaker 1 vs. Multi 2 → Speaker 2), *exposure group*, *target type*, *prime type* and their interactions as fixed effects. Priming effects are shown in Fig.6, presented separately by test talker. Crucially, there was a significant four-way *talker condition* × *target type* × *prime type* × *exposure group* interaction ($\beta = -5.63$, SE = 2.05, $p < .01$), suggesting that the priming pattern was affected by the specific exposure-test speaker pair. Additional mixed-effects models were fitted for each talker condition separately. In each model, *exposure group*, *target type* and *prime type* as well as their interactions were fixed effects. The structure of random effects remained the same. For Multi 1 → Speaker 1 condition, there was a significant three-way *exposure group* × *target type* × *prime type* interaction ($\beta = -6.73$, SE = 2.87, $p < .05$). Follow-up analyses for -/d/ words revealed a significant priming effect ($\beta = -23.85$, SE = 4.59, $p < .001$), modulated by a significant

exposure group × *prime type* interaction (β = 9.87, SE = 4.38, $p < .05$). By-group analysis further indicated a significant priming effect for -/d/ words in the experimental group (β = -32.93, SE = 6.70, $p < .0001$), but not in the control group (β = -11.38, SE = 6.41, $p = .08$). For /t/-final words, there was a significant priming effect (β = -11.65, SE = 4.61, $p < .05$) but no interaction between *exposure group* and *prime type* (β = -3.46, SE = 3.74, $p = .35$), suggesting equivalent priming magnitudes between groups. This result was highly comparable to effects we have found for talker-specific adaptation in Experiment 1.

However, for Multi 2 → Speaker 2 condition, despite a significant priming effect (β = -15.97, SE = 3.19, $p < .0001$), there was no three-way *exposure group* × *target type* × *prime type* interaction (β = 4.11, SE = 2.82, $p = .14$). No other effects were significant at the .05 level. Thus, exposure to multiple talkers seemed to have enhanced word recognition for words produced by speaker 1 but somehow not for words produced by speaker 2.

In order to understand why listeners generalized to speaker 1 only, we analyzed participants' answers to the accent similarity questions. When asked to identify the number of talkers during exposure and test, all listeners indicated that they heard multiple talkers during exposure (answers ranging from two to seven). Interestingly, 28 out of 48 participants thought there was more than one talker at test too. When asked to guess the type of accents, only three participants (one from the experimental group and two from the control group) identified all talkers as having a Mandarin accent; other listeners were generally not confident and their answers included all kinds of accents (e.g., Asian, Spanish, Middle East, European and native American), suggesting they were unfamiliar with the Mandarin accent. Of particular interest is whether listeners perceived the exposure talkers and test talkers to be similar. Although we asked participants to indicate whether the talkers had the same accent or different accents, many

participants indicated that they were similar but different; and some expressed low confidence about their answers. We thus asked everyone to give a rating on a scale of 1-7 (regardless whether they identified the same accent or not). In Multi 1 → Speaker 1 condition, eight participants reported "different accents", ten participants reported "similar accents", and six participants reported "same accent"; in Multi 2 → Speaker 2 condition, ten participants reported "different accents", thirteen participants reported "similar accents", and only one participant identified them as having the same accent. Likert ratings of accent similarity on a scale of 1-7 yielded more gradient results. Speaker 1 as the test talker (M = 4.85, SD = 1.16) was rated as more similar to exposure talkers as a group than speaker 2 (M = 4.50, SD = 1.41), although Mann-Whitney U test indicated that the difference between speakers was not significant (p = .34). The subjective reports on accent similarity suggested two things: on the one hand, overall listeners did not perceive the exposure and test talkers to have the same accent, thus it was unlikely that listeners generalized to speaker 1 due to an active strategy to extract commonalities across talkers; on the other hand, speaker 1 was indeed perceptually more similar to the other talkers than speaker 2. It is possible that listeners developed some tacit knowledge about the acoustic properties of /d/ following exposure to a group of talkers, and such knowledge was more applicable to the productions of speaker 1 due to acoustic similarity.

Analyses were conducted on three acoustic cues of critical /d/ words to see if acoustic similarity could provide a more clear account of the differential generalization pattern. As noted in Experiment 4, exposure words had more syllables than test words and it was not fair to compare temporal cues (duration of preceding vowel, closure and burst) between exposure words and test words. Because test words were only available for speaker 1 and speaker 2, I reasoned that a comparison between the exposure words produced by exposure talkers as a group and

exposure words by speaker 1 and 2 (used in Experiment 1) would help us to gauge the degree of similarity between the exposure and test talkers here. I compared the production patterns of exposure words between exposure talkers as a group (Multi 1 and Multi 2) and the test talkers (Speaker1 and Speaker 2), separately for each exposure-test pair. Independent samples t-tests indicated that Speaker 1 produced critical words with significantly shorter closure than speakers in Multi 1 group did ($t(58) = 4.169$, $p < .001$), but they had similar patterns for vowel ($t(58) = 1.145$, $p = .26$) and burst ($t(58) = .11$, $p = .91$). In contrast, Speaker 2 had significantly longer duration for all three acoustic cues than speakers in Multi 2: vowels ($t(58) = 3.755$, $p < .001$), closures ($t(58) = 2.789$, $p < .001$), bursts ($t(58) = 5.877$, $p < .001$). Thus, the acoustic measures paralleled the explicit ratings given by the participants: Speaker 1 was indeed more acoustically similar to the exposure talkers as a group[4].

Discussion

Our results indicated that multiple-talker exposure can allow listeners to retune the sound-to-category mapping for word-final /d/ in a talker-independent way; and when it occurs, the phonetic retuning leads to improved word recognition for a novel talker. However, such cross-talker generalization was constrained by the inter-talker similarity in the productions of the critical segment. The results provided some evidence that despite a shared native language (L1), non-native speakers may differ in the realizations of some phoneme contrasts in L2; such inter-

4 As an exploratory analysis, we examined whether speaker 1 aligned with any of the five exposure speakers in particular. Pairwise comparisons were conducted to compare exposure words produced by speaker 1 to those produced by each exposure speaker. The only significant differences were that speaker 1 had longer bursts than speaker 3 ($p < .01$) and shorter bursts than speaker 5 ($p < .05$); he also had shorter closures than speaker 6 ($p < .05$). Of interest, speaker 1 had remarkably similar patterns to speaker 4, with no difference between them on any of the three acoustic measures (ps > .50). In contrast, speaker 2 differed from every exposure speaker in Multi 2 group on at least two out of the three measures, significantly at the .05 level.

talker misalignment in the distribution of acoustic properties may hinder generalization. Talker dissimilarity may explain why in some cases, training with multiple talkers of the same accent improved speech intelligibility of trained talkers only but did not generalize to novel talkers (Wade et al., 2007; Clarke, 2000), whereas other evidence shows generalization from a set of talkers to a new talker (e.g., Bradlow & Bent 2008).

Of note, a pilot intelligibility study showed that the exposure speakers as a group were of equivalent intelligibility to the test speakers and all speakers produced /t/-like /d/ words to some extent. The lack of generalization from five Mandarin-accented speakers to Speaker 2 implies that listeners did not merely perceive the Mandarin-accented speakers as people who "produced /d/ like /t/s". Rather, they were sensitive to the fine-grained phonetic detail in foreign-accented speech. Listeners' subjective reports of accent similarity were consistent with acoustic measures. Acoustic analyses revealed that the temporal patterns of acoustic cues in Speaker 2's productions were different from other exposure talkers as a group and as individuals. In Experiment 1, native-English listeners adapted to Speaker 1 and showed a re-weighting of acoustic cues, favoring burst length over vocalic cues as an informative cue to Mandarin-accented voicing tokens. To speculate, listeners might have engaged in the same kind of perceptual adjustments for the exposure talkers in the current experiment. When perceiving misalignment along specific acoustic dimensions, they may implicitly perceive the tokens from Speaker 2 as dissimilar to the exposure talkers and therefore do not generalize from exposure to test. Cross-talker generalization appears to be an implicit process because both Experiment 4 and Experiment 5A showed that listeners' explicit judgment of the similarity voices or accents did not allow listeners to generalize when the acoustic information was dissimilar (Experiment 4), nor did a judgment of accent dissimilarity extinguish generalization when acoustic information supported it. In

Experiment 5A, listeners' reports indicated poor awareness of a shared accent among talkers, and generalization was found in the Multi 1 → Speaker1 condition, regardless of whether individual participants had identified the talkers' accent as "same" or "different".

Experiment 5B

Experiment 5A showed that listeners learned about the acoustic-phonetic structure of /d/ tokens in mixed speech coming from five Mandarin-accented speakers (Multi 1) and generalized the learning to a novel Mandarin speaker (Speaker 1). In Experiment 1, talker-specific perceptual learning was found for Speaker 1 (Speaker 1 → Speaker 1). Of interest is whether the perceptual benefits originated from multiple-talker exposure in Experiment 5A was comparable to talker-specific exposure. In Experiment 1, twenty-four monolingual English participants were included for each exposure group (experimental vs. control). Comparison of the two studies necessitated the addition of more participants in the Multi→Speaker1 condition to match the sample size with Speaker 1 → Speaker 1 condition in Experiment 1 in order to assess the effects of learning from a specific talker versus learning from multiple talkers.

Methods

Participants. Twenty-four additional participants participated in the experiment. Including those participants from Experiment 5A who were tested on speaker 1 (n = 24), a total of forty-eight participants were included in the following analyses, with equal numbers of participants in the experimental and the control condition.

Materials and procedure. All materials and procedures were identical to those of the Multi 1 → Speaker 1 condition in Experiment 5A.

Results

Exposure. Response accuracy for critical /d/ words in the exposure group (*M* = .79, *SD* = .09) suggests that they were judged to be real words for the most part (see Table C1).

Test. Table C3 shows mean error rates and RTs in the test phase. Priming effects are shown in Fig.7 (left panel). Responses (4.8% of correct trials) above or below 2 SDs from the mean of each prime type in each exposure group were excluded from the RT analysis. The same mixed-effects model analyses were conducted as in Experiment 1. There was a significant priming effect (β = -15.83, SE = 2.87, $p < .0001$). /d/-final targets elicited slower responses than /t/-final targets (β = 14.93, SE = 5.82, $p < .05$). There was a significant main effect of *exposure group* (β = -20.30, SE = 2.47, $p < .001$), driven by overall faster responses in the experimental group than the control group[5]. Of interest, there was a significant three-way *exposure group* × *target type* × *prime type* interaction (β = -6.01, SE = 2.23, $p < .01$). No other effects were significant at the .05 level.

Follow-up analyses were conducted by fitting mixed-effects models for /d/-final targets and /t/-final targets separately, with *exposure group*, *prime type*, and their interaction as fixed effects. Crucially, for /d/-final targets, the priming effect was significantly larger in the experimental group than the control group, as revealed by an *exposure group*-by-*prime type* interaction (β = -9.17, SE = 3.40, $p < .001$). For /t/-final targets, there was no *exposure group*-by-*prime type* interaction (β = 2.87, SE = 2.93, $p = .33$). I also asked whether within each exposure group, the priming magnitudes would differ between target types. Starting with the control group, there was a main priming effect (β = -12.29, SE = 3.77, $p < .01$) but no interaction between *target type* and *prime type* (β = 5.59, SE = 3.77, $p = .14$), suggesting that auditory –/d/ words primed -/d/ and -/t/ targets equally. In contrast, for the experimental group, a main priming effect

[5] Given that analysis on filler items also showed similar group effect, this group difference is likely due to between-subject variability.

(β = -19.08, SE = 3.13, $p < .001$) was modulated by a *prime type*-by-*target type* interaction (β = -7.92, SE = 3.14, $p < .05$), driven by larger priming for "*seed – SEED*" type trials than for "*seed – SEAT*" types. Follow-up tests indicated that priming effects were significant for both /d/-final targets (β = -28.07, SE = 4.88, $p < .0001$) and /t/-final targets (β = -10.76, SE = 4.07, $p < .05$) within the experimental group.

Across-experiments analysis. The perceptual learning effects as exhibited by group difference in priming patterns replicated the previous finding on talker-specific learning (Speaker 1 → Speaker 1, Experiment 1): prior exposure to critical /d/ words significantly increased the degree of match between the auditory signal of other /d/-final words and their word forms (e.g., seed) in the mental lexicon, making *seat*-like words a weaker lexical competitor to *seed*-like words and facilitated word recognition. To statistically assess whether learning resulted from generalization was as good as learning from the specific talker, I pooled data across studies and included *experiment* (Experiment 1: Speaker1 → Speaker 1, coded as 1, and Experiment 5B: Multi 1 → Speaker 1, coded as -1) as an independent variable into the mixed-effects model. Fixed effects included *experiment*, *exposure group*, *target type*, *prime type* and full-scale interactions between these factors. A maximal random-effects structure justified by the data was used. Results revealed a priming effect (β = -19.63, SE = 2.14, $p < .001$). There was a significant *prime type*-by-*experiment* interaction (β = -3.52, SE = 1.52, $p < .05$), with the priming effects being smaller overall in the current experiment. There was a main effect of *target type* (β = 14.99, SE = 4.93, $p < .01$). There was a significant three-way *exposure group* × *target type* × *prime type* interaction (β = -4.93, SE = 1.62, $p < .01$), reflecting a larger priming for /d/-final targets than /t/-final targets in the experimental group only, as in Experiments 1 and 5B. However, this three-way interaction term did not further interact with *experiment* (β = .95, SE =

1.61, p = .56), indicating that the generalization pattern did not differ significantly across experiments.

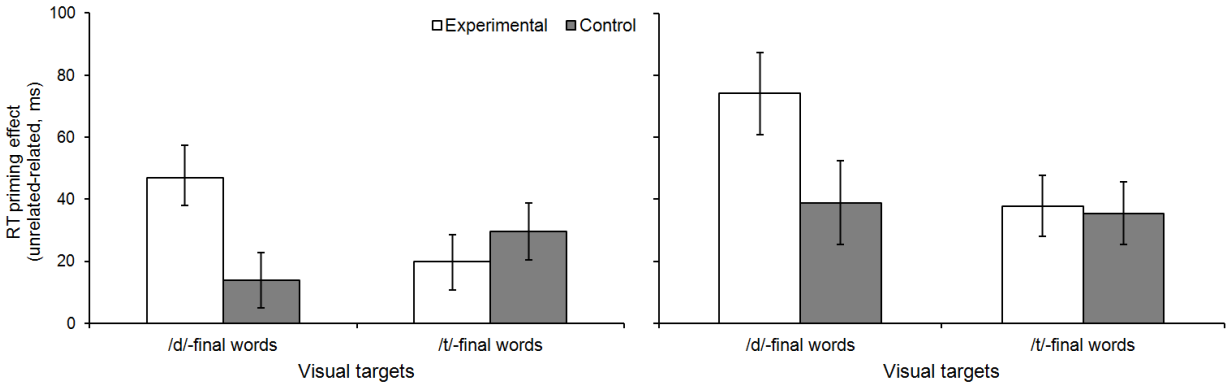

Fig.7. Test phase results across studies: multiple-talker condition (Multi 1 → Speaker 1, Experiment 5B, left panel) and talker-specific condition (Speaker 1 → Speaker 1, Experiment 1, right panel). Priming of /d/-final words (e.g., *seed*) and /t/-final words (e.g., *seat*) as a function of exposure group (Experimental versus Control). In both related priming types, /d/-final words (e.g. *seed*) served as auditory primes. Error bars represent standard errors of the mean.

Discussion

We conclude that multiple-talker exposure was as effective as talker-specific exposure in helping exposure participants gain an advantage over control participants and in attenuating lexical competition between /d/- and /t/-final minimal pairs. Bradlow and Bent (2008) trained participants with sentence-level speech stimuli and established a learning effect from multiple-talker exposure as large as that from talker-specific exposure. In their study, the learning effect was defined as an increase in word-level recognition accuracy, measured by the number of keywords in sentences. We replicated their finding at the phoneme level. Note, however, despite that experimental participants exhibited an advantage over control participants following multiple-talker exposure within Experiment 5B, the overall priming was weaker following multiple-talker exposure (relative to talker-specific exposure). In this regard, our results also suggest that learning from a specific talker gave listeners the most benefit in word recognition.

Across experiments presented in this chapter, there were two major findings. First, talker-specific phonetic retuning was not readily transferrable to a novel talker of the same accent. Exposure to a group of talkers (instead of a single talker) increased the likelihood that listeners could learn the accent and generalize learning to a novel accent. Second, in both cases, explicit knowledge of talker identity or talker accents was not the decisive factor constraining generalization across talkers. Rather, bottom-up similarity between exposure and test talkers had direct consequences on talker generalization. There are three questions we may ask about our results, situated in the findings of past research: Why do listeners appear to generalize experience from one speaker to another in some cases whereas sometimes they do not? Is there an interplay between top-down information about the talker situation (e.g., who is speaking, how many talkers, what kind of accents) and bottom-up acoustic information in guiding talker generalization? How do listeners move from talker-specific adaptation to general accent adaptation? We discuss our answers to these questions in turn.

Why do listeners appear to generalize experience from one speaker to another in some cases? Previous research has yielded divided patterns with respect to talker generalization: for fricatives, adjustments seem to be talker-specific (Kraljic & Samuel, 2005, 2007; Eisner & McQueen, 2005) and listeners maintain separate sound-to-category mappings for different speakers. In comparison, talker adaptation for stop consonants exhibits exactly the opposite pattern: a single sound-to-category mapping was employed and phonetic adjustments were carried over to different speakers (Kraljic & Samuel, 2005, 2007). Our findings on stop consonants together with the findings of Reinisch and Holt (2014) on fricatives make it clear that the different generalization patterns for stops versus fricatives are a by-product of bottom-up similarity in the segmental productions. For both types of phonemes, listeners do not generalize

to novel talkers if the production pattern of specific phonemes from the new speaker does not match their experience from a prior speaker; furthermore, they will readily generalize to a different talker if bottom-up similarity supports it. Note that in Reinisch and Holt (2014), bottom-up similarity was measured by listeners' perception of the degree of ambiguity of the critical segments. Our results further revealed that listeners were not merely assessing speaker similarity based on their overall ambiguity (as speakers were matched on this measure in Experiment 4), they were sensitive to fine-grained variation along multiple acoustic dimensions and a comparison of talker' acoustic-phonetic space to prior talkers seemed to constrain the interpretation of linguistic categories in the talker's productions.

This leads us to the second question: does acoustic similarity tell the whole story or is there a role of top-down influence from listeners' perception of the talker situation? Eisner and McQueen (2005) cross-spliced ambiguous fricative sounds produced by one speaker into an entirely new voice and observed a typical adaptation pattern for the ambiguous sounds, despite the fact that the new voice was perceptibly different. That is, the context of speech (or perceived voice) in which the critical segment was embedded did not matter. Similarly, Reinisch and Holt (2014) found that listeners generalized their experience with a prior accented speaker to a novel speaker, despite that the two speakers were identified as two individuals of different accents. In Experiment 4, we did not find evidence of generalization even among participants who believed they were listening to a single speaker the whole time. These findings mark the significance of bottom-up similarity in governing generalization across talkers. However, this is not to say that similarity between old and new speech stimuli is the sole reason whether listeners generalize or not. Evidently, top-down knowledge of talker identity does make a difference in phonetic retuning. Samuel and Kraljic (2013) manipulated listeners' expectations of talker identity by

presenting visual information of a single speaker or two different speakers. While the same speech tokens were used, phonetic retuning for fricatives differed as a function of listeners' expectations: a mix of ambiguous and unambiguous tokens denoting the same phoneme blocked phonetic retuning only when listeners expected them to be spoken by the same voice. Together, the findings are consistent with a framework in which listeners build up conservative models to represent talker-specific phonetic categories. Generalization is observed only when talkers are sufficiently similar. This raises the question: will listeners develop more talker-general representations as they gain more experience with a particular accent?

We thus arrive at the third question: How do listeners move from talker-specific adaptation to general accent adaptation? Previous work has shown that exposure to multiple talkers is beneficial for accent adaptation, namely, adaptation for the specific accent independent of individual speakers (Bradlow & Bent, 2008; Sidaras et al., 2009). Researchers have compared talker-independent adaptation to foreign accents to findings from the literature on perceptual learning for speech. Results of Experiments 5A and 5B provided the first direct evidence that exposure to multiple talkers indeed elicited phonetic retuning that was generalizable within the accent to a novel talker. We point out that, although it was often hypothesized that distillation of some systematic phonetically-relevant properties across talkers led to talker-independent adaptation, our data suggested that "systematic properties" may not be demonstrated by all talkers of the accent: inter-talker variability is large in non-native speakers. Our results also indicated that listeners were sensitive to bottom-up similarity among talkers and used it to guide talker generalization. Given this, we suggest another theoretically quite different possibility that might also explain the current and previous findings on accent adaptation: multiple-talker exposure provides a larger exemplar pool (a larger sampling of acoustic-phonetic space) to

which novel talkers can be compared and increases the probability of encountering a similar talker. It is not necessary that all exposure talkers demonstrate systematicity in their productions and it does not require listeners to be explicitly aware of a shared accent; it requires one or more talkers to be sufficiently similar to the test talker. If this is the case, then in essence, multiple-talker exposure benefits generalization as well as an appropriate exemplar speaker does in a single talker exposure. Of interest, an exploratory analysis (see footnote 4) did reveal that in the Multi 1 → Speaker 1 condition where we observed generalization, the test talker (speaker 1) was aligned with one of the exposure talkers (speaker 4) on every acoustic measure. Thus, with speaker 1, listeners may not only find him similar to the acoustic properties of the exposure talkers in aggregate, they could also latch onto exposure speaker 4 as a comparable exemplar. The current experiments are unable to differentiate between the *extraction* hypothesis and the *exemplar* hypothesis. Future studies should test whether one "close-enough" exposure talker, among a set of very dissimilar talkers (different accents, for instance), would enable generalization to an acoustically-similar test talker, even when the test talker does not share any commonality with other talkers. In brief, we suggest that *acoustic similarity* between talkers explains why listeners sometimes fail to generalize across talkers and why they can benefit more from multiple talker exposure.

CHAPTER 5 THE LIMITS OF PHONETIC ADAPTATION

In this chapter, I present two experiments that serve as case studies of talker-specific adaptation to foreign-accented speech. Previous research has reported successful phoneme-level adjustments in the face of acoustic-phonetic variation (e.g., Reinish & Holt, 2014; Eisner et al., 2013; Witteman et al., 2013). No study has addressed the effects of speaker intelligibility and intra-talker variability in phonetic retuning, despite some evidence showing these factors have direct influences on the transcription accuracy of accented words (Wade et al., 2007; Bradlow & Bent, 2008). Specifically, I aim to tease the two factors apart and investigate their independent contribution to successful adaptation to acoustic-phonetic variation in foreign-accented speech. I adopted the same experimental paradigm as used in Experiment 1 and investigated talker-specific adaptation to two different Mandarin-accented speakers. The two speakers were selected such that one speaker had comparable baseline intelligibility to Speaker 1 but exhibited larger within-category variability in the production of critical phonemes (Speaker HV, short for 'high variability', used in Experiment 6); another speaker had lower intelligibility than the original speaker but demonstrated very small intra-talker variability in his productions (Speaker LI, short for 'low intelligibility', used in Experiment 7). The mean intelligibility for critical test /d/-final words was comparable between speaker 1 (M = .65, SD = .13) and speaker HV (M = .70, SD = .06), with no significant difference, $t(18) = 1.061$, p = .30. Speaker LI (M = .22, SD = .17) had much lower intelligibility for /d/ tokens than Speaker 1 and HV (see Appendix A for detail). Standard deviation and range of the three acoustic measures (vowel duration, closure length, and burst length) was taken as the measures for acoustic variability (see Table 4). The measures for Speaker 1 are also presented for comparison. A comparison of adaptation results for these three speakers would illustrate whether within-talker variability and speaker intelligibility are

independently linked to the ease of foreign accent adaptation. More broadly, the results will shed light on the constraining factors of perceptual learning and are pertinent to the development of computational models for mechanisms that allow phonetic representation adjusted on a talker-by-talker basis.

Table 4. Three acoustic cues (preceding vowel duration, closure interval duration, and length of burst and aspiration) were measured for each word, presented in msec. Standard deviations and range of each temporal cue across words are reported as measures of within-talker variability.

Speaker	Standard deviations			Range		
	Vowel	Closure	Burst	Vowel	Closure	Burst
Exposure /d/-final words						
Speaker HV	46	19	30	164	83	111
Speaker LI	31	18	14	119	84	63
Speaker 1	34	18	12	150	92	46
Test /d/-final words						
Speaker HV	38	16	23	194	64	112
Speaker LI	23	14	7	117	92	34
Speaker 1	34	16	13	142	90	57

Experiment 6

Methods

Participants. Forty-eight monolingual English speakers with no hearing or visual problems (according to self-report) were recruited from the University of Connecticut community. All participants were undergraduate students who were naïve to Mandarin and had no or minimal previous exposure to Mandarin-accented English. Participants were randomly assigned to one of the two exposure groups (experimental vs. control). After excluding one participant for misunderstanding the test task, twenty-three experimental participants and twenty-

four control participants were included in the following analyses.

Speech materials. Speaker HV was Speaker 2 in Experiments 4 and 5. We use "HV" here to indicate his speech characteristics of interest in this experiment. Speech materials and recording procedure was identical to that in Experiment 1.

Procedure. The experimental procedure was identical to that in Experiment 1. Talker-specific perceptual learning was examined.

Results

Exposure. For Speaker HV, response accuracy indicated that critical /d/ words were largely judged to be real words by the exposure group (M = .82, SD = .15). Accuracy for other words is presented in Table C1.

Test. Table C4 shows mean error rates and RTs in the test phase. RT priming effects are shown in Fig.8 (left panel). Responses (4.6% of correct trials) above or below 2 SDs from the mean of each prime type in each exposure group were excluded. A mixed-effects model was fitted as in Experiment 1. There was a main effect of *prime type* (β = -12.27, SE = 2.80, p < .0001) and a main effect of *target type* (β = 13.99, SE = 5.34, p < .05). There was an interaction between *exposure group* and *target type* (β = 4.55, SE = 2.12, p < .05). Crucially, there was no three-way interaction between *target type*, *prime type* and *exposure group* (β = -.34, SE = 2.12, p = .87). So there was no evidence of adaptation in the experimental group, compared to the control group.

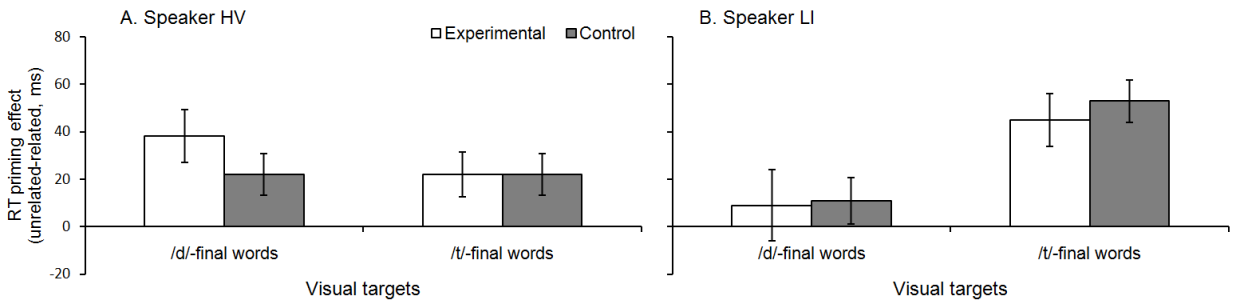

Fig.8. Test phase results for Speaker HV (panel A, Experiment 6) and Speaker LI (panel B, Experiment 7): Priming effect (RT in unrelated priming minus RT in related priming) to /d/-final words (e.g., *seed*) and /t/-final words (e.g., *seat*) for participants exposed to critical words (experimental group) or replacement words (control group). In both related priming types, /d/-final words (e.g., *seed*) served as auditory primes. Error bars represent standard errors of the mean.

Comparing Speaker HV with Speaker 1. To statistically assess whether learning resulted indeed differed between the two speakers, I pooled data across studies and included *experiment* (Experiment 6: Speaker HV → Speaker HV and Experiment 1: Speaker1 → Speaker 1) as an independent variable (contrast coded as follows: Experiment 1 = 1, Experiment 6 = -1). Fixed effects included *experiment*, *exposure group*, *target type*, *prime type* and full-scale interactions between these factors. A maximal random-effects structure justified by the data was used. Results revealed a priming effect (β = -18.13, SE = 2.10, $p < .001$). There was a significant *prime type*-by-*experiment* interaction (β = -5.29, SE = 1.50, $p < .001$), with the priming effects being smaller overall in the current experiment. There was a main effect of *target type* (β = 14.27, SE = 4.90, $p < .01$). Unlike in Experiment 1 alone, there was now no three-way *exposure group* × *target type* × *prime type* interaction across experiments (β = -2.14, SE = 1.50, $p = .15$). However, the four-way interaction between the four independent variables was not significant either (β = -1.90, SE = 1.50, $p = .20$). Therefore, although a significant adaptation effect was not

observed between experiment group and control group for Speaker HV, the overall priming patterns did not differ for Speaker HV and Speaker 1 significantly.

Discussion

Given that Speaker HV had equivalent intelligibility to that of Speaker 1 in Experiment 1, successful talker-specific adaptation in the previous study and the lack of significant adaptation effects in the current study together pointed to a negative effect of intra-talker variability on phonetic retuning. However, /d/-final words produced by Speaker HV did elicit numerically greater activation for the intended targets among the experimental group than the control group, suggesting that the intra-talker variability may have slowed down the adaptation processing, or dampened adaptation effects in some participants, instead of blocking it. A lack of significant difference across Experiment 1 and Experiment 6 further supports this. However, given the null interaction effects involving "experiment," additional studies are needed to examine whether more exposure tokens would ultimately help listeners to retune phonetic representation of the /d/ category when intra-talker variability is large.

A question stemming from this finding is that whether listeners would have difficulty adapting when the overall stimulus variability is large in the speech signal, regardless of individual speakers. Given the successful generalization in Experiment 5B in the condition of multi-talker exposure (which likely increased the overall variability), this was unlikely. A comparison between the Multi 1 → Speaker 1 in Experiment 5B and Speaker HV → Speaker HV in this experiment confirmed this. On the variability measures, the critical exposure words from Multi 1 as a group had a standard deviation of 48ms (vowel), 27ms (closure) and 33ms (burst) and a range of 180ms (vowel), 106ms (closure) and 146 ms (burst). All of these measures were larger than those of the productions of Speaker HV, indicating larger variation in Multi 1

group. However, it did not keep listeners from generalizing to Speaker 1. Thus, large inter-talker variability *per se* did not nullify adaptation. The current results suggest that difficulty with Speaker HV reflects listeners' ability to track *talker-specific* acoustic patterns in adaptation.

Experiment 7

Methods

Participants. Forty-four participants listened to Speaker LI, with twenty-three participants in the experimental group and twenty-one participants in the control group. All participants gave informed consent and received course credits for their participation.

Speech materials. Speaker LI was a late L2 learner of English. Demographic information of speaker is presented in Appendix A. Speech materials and recording procedure was identical to that in Experiment 1.

Procedure. The experimental procedure was identical to that in Experiment 1.

Results

Exposure. For Speaker LI, fewer of the critical words were judged to be real words by the experimental group (*M1* = .61, *SD1* = .10), relative to Speaker HV (in Experiment 6) and Speaker 1 (in Experiment 1). Replacement words were judged with much higher accuracy by the control group (*M2* = .81, *SD2* = .06), which was comparable to that for the other two speakers. Response accuracies for filler words and nonwords were not different between the two exposure groups (See Table C1). This suggests that the overall low accuracy in lexical decision for LI is mostly attributable to consistent misperception of /d/-final tokens as /t/ (e.g. 'overload' is perceived as 'overloat'). The low accuracy for critical words was expected for Speaker LI, as he was selected based on his low intelligibility productions of /d/ tokens. The question was whether

the small amount of words that were recognized as real words were sufficient to provide lexical-to-phonetic feedback that drives phonetic retuning. Of note, Kraljic and Samuel (2007) showed that as few as ten critical ambiguous items were enough to elicit a recalibration of the boundary between two contrastive phonetic categories (e.g., /d/ and /t/).

Test. Table C4 shows mean error rates and RTs in the test phase. RT priming effects are shown in Fig.8 (presented together with Speaker HV in Experiment 6). Responses (5.0% of correct trials) above or below 2 SDs from the mean of each prime type in each exposure group were excluded. A mixed-effects model was fitted as in Experiment 6. There was a main effect of *prime type* (β = -15.47, SE = 2.88, $p < .0001$) and a main effect of *target type* (β = 26.96, SE = 4.97, $p < .0001$). There was an interaction between *target type* and *prime type* (β = 9.91, SE = 2.88, $p < .001$). Crucially, there was no three-way interaction between *target type*, *prime type* and *exposure group* (β = -.22, SE = 2.48, $p = .93$).

To unpack the *target type* × *prime type* interaction, additional mixed-effects models were fitted separately for /d/-final targets and /t/-final targets. Results indicated a significant priming effect only for /t/-final targets (β = -25.68, SE = 3.96, $p < .0001$), but not for /d/-final targets (β = -5.37, SE = 4.20, $p = .21$). That is, auditory forms of "seed" successfully primed the visual targets of "SEAT" instead of "SEED". Furthermore, there was no interaction between *exposure group* and *prime type* for either target type (ps > .80). This significant target type effect in priming magnitude and a lack of further interaction with *exposure group* suggests that, regardless of exposure group, when listeners heard the accented /d/-final words (e.g., 'seed'), /t/-final word forms (e.g., 'seat') were activated to a greater extent in the mental lexicon than the intended words. This was expected for the control group in that the low intelligible /d/ tokens should pose a mismatch to the mental representation of /d/ and be mapped onto /t/ category. However,

surprisingly, there was no sign of any adaptation benefit observed in the experimental group. Given that acoustical variability of Speaker LI was comparable to Speaker 1 in Experiment 1 (Table 4, with Speaker LI exhibiting even higher within-talker consistency than Speaker 1), the difference in priming patterns were likely due to the low intelligibility of Speaker LI. This implies that sufficiently intelligible tokens are a pre-requisite for rapid acoustic-phonetic adaptation.

Comparing Speaker LI with Speaker 1. Again, to statistically assess whether learning resulted indeed differed between the two speakers, we pooled data across studies and included *experiment* (Experiment 7: Speaker LI → Speaker LI and Experiment 1: Speaker1 → Speaker 1) as an independent variable (contrast coded as follows: Experiment 1 = 1, Experiment 7 = -1). Fixed effects included *experiment*, *exposure group*, *target type*, *prime type* and full-scale interactions between these factors. A maximal random-effects structure justified by the data was used. Results revealed a priming effect ($\beta = -19.78$, SE = 1.98, $p < .001$) and a significant *prime type*-by-*experiment* interaction ($\beta = -3.82$, SE = 1.62, $p < .05$), with the priming effects being smaller overall in the current experiment. There was a main effect of *target type* ($\beta = 21.09$, SE = 4.80, $p < .001$). In addition, there was a significant *target type* × *experiment* interaction ($\beta = -6.40$, SE = 1.72, $p < .001$) and a significant *target type* × *prime type* × *experiment* interaction ($\beta = -7.11$, SE = 1.62, $p < .001$), suggesting that the priming patterns for the two target types differed between the two experiments. This was consistent with separate analyses conducted for the two experiments. No other effects were significant. To unpack the interaction *target type* × *prime type* × *experiment* interaction, two additional models were fitted, separately for /d/-final words and /t/-final words. In each model, fixed effects included *experiment*, *exposure group*, *prime type* and their interactions. For /d/-final targets, there was a significant effect of

experiment-by-prime type interaction (β = -10.73, SE = 2.56, $p < .001$), reflecting the fact that identity priming for /d/-final words was larger overall (across experimental and control groups) following talker-specific learning for Speaker 1 in Experiment 1 than for Speaker LI. This was due to the extremely small priming for Speaker LI's /d/ productions. For /t/-final words, the priming magnitude did not differ as a function of *exposure group* or *experiment* (*p*s > .10).

Discussion

Previous literature establishes that native listeners flexibly accommodate unfamiliar acoustic-phonetic variation in speech via a mechanism of phonetic retuning guided by top-down knowledge such as lexical information (e.g., Norris et al., 2003; Eisner et al., 2013). Phonetic retuning has been shown to happen over a very short time scale: a few critical words containing the to-be-adapted segment are sufficient to drive perceptual learning. Little is known about the limits of such perceptual learning. To our knowledge, the present work is the first study to specifically address the role of speaker intelligibility and intra-talker variability in phoneme-level adaptation. Our results showed that both of them independently had direct consequences for rapid perceptual adaptation. Low speech intelligibility likely sets limits on the amount of top-down lexical feedback to phonetic adjustments, although we could not entirely exclude the possibility that lower intelligibility talkers may have larger phonetic deviations from the target phonemes, which may require more time to adjust to. However, this possibility itself is untested. We return to this in the general discussion in Chapter 6. Large intra-talker variability may have increased the degree of indeterminacy of mapping bottom-up acoustic signal onto speech categories thus slowing down adaptation, as shown in the comparison of adaptation for Speaker 1 and Speaker HV. It is important to note that the negative impacts did not affect every listener. Rather, it increased variability of adaptation effects among participants. As the across-

experiments analyses show, although there was no significant adaptation effect within Experiment 6, whereas there were significant effects of adaptation in Experiment 1, group differences in priming patterns were not significantly different between these two experiments. It is possible some participants were more susceptible to negative influences coming from acoustic variability.

At least, we could say that high intelligibility alone is not a guarantee of adaptation success. Future studies should further investigate whether the observed negative influences on adaptation merely reflect initial delays and can be overcome by more speech input, as shown with adaptation to sentence-level stimuli (Bradlow & Bent, 2008), or alternatively, represent more categorical limitations on phonetic retuning. For instance, it might be that failing to meet an intelligibility threshold blocks rapid adaptation completely and only extensive training can improve perception of poorly intelligible accented speech (see Wade et al., 2007).

Native phonetic retuning (e.g., Kraljic & Samuel) is largely thought to be automatic and rapid; in contrast, perceptual learning for non-native speech categories generally requires long and repetitive training (e.g., Lively, Logan, & Pisoni, 1993; Bradlow, 2008). The different timelines for these two types of learning suggest distinct perceptual and cognitive processes. It is an open question whether as foreign-accented speech becomes more non-native like and more variable, it would evoke the adaptation mechanism needed for non-native category acquisition and make the adaptation cognitively more taxing. Notwithstanding, the current results mark the difficulty listeners may face in adapting to natural foreign accents.

CHAPTER 6 GENERAL DICUSSION

Speech perception requires listeners to extract a meaningful message out of highly variable and sometimes ambiguous signals. One prominent source of speech variability arises from talker-related characteristics. Despite considerable talker variation, past research has shown that listeners are very adept at accommodating changes in speaking rate (Miller & Liberman, 1979), idiosyncratic pronunciations (e.g., Norris et al., 2003), an unfamiliar dialect (e.g., Floccia, Goslin, Girard & Konopczynski, 2006) or foreign accent (e.g., Clarke & Garrett, 2004). The exact mechanism by which adult listeners efficiently accommodate talker variability in natural speech is not well understood. Nor is there a consensus with respect to the nature of lexical and pre-lexical representations. Research on 'perceptual learning for speech' reveals that native listeners use top-down information to flexibly adjust the mapping from speech sounds to phonetic categories in the face of perceptual ambiguity (e.g., Norris et al., 2003; Kraljic & Samuel, 2005, 2006, 2007). In this dissertation, I adopted the exposure-and-test perceptual learning paradigm (Norris et al., 2003; McQueen et al., 2006) to study native listeners' rapid adaptation to a special case of talker variation, namely foreign accents.

Perceiving foreign-accented speech is a particularly challenging task, because it not only contains typical idiolectal differences, but also presents more global deviations from native language categories. Native listeners may have to reorganize their own representations of phonetic categories because accented speech does not map well onto existing inventories. Furthermore, inevitable influences of phonological transfer from one's native language can make a non-native speaker's phonetic categories unstable and likely increase within-talker variability (Wade et al., 2007). Needless to say, speakers differ in their L2 proficiency and speaker intelligibility can vary considerably across speakers of the same accent. Meanwhile, exactly due

to systematic influences from their L1, speakers with the same non-native accent do share some accent regularities in their speech. Given this, there are potentially strong motivations to generalize across non-native speakers of the same accent whereas generalizing across idiolect differences has less utility. Together, these characteristics make foreign-accented speech a good experimental case to examine how listeners form talker-specific representations during adaptation and how the representations may evolve and generalize across talkers. I now summarize my findings and situate them in the broader literature on speech perception and adaptation for a discussion of insights we may have from the current study.

What is Reorganized during Talker-Specific Adaptation?

While previous work has focused on the recalibration of phonetic boundaries, my *first* major finding is that the learning of systematic variation in accented speech fundamentally affects the phonetic analysis of speech samples: even *unambiguous* tokens were perceived as better exemplars of the intended categories (/d/ and /t/) after exposure to an accented speaker. Thus, listeners were not merely moving the phonetic boundary in order to resolve lexical ambiguity; in addition, they were sensitively tracking the fine-grained phonetic detail that carries critical segmental information. Adaptation to the accented speech reshaped the internal phonetic structure beyond the boundary region. This finding adds to existing literature that characterizes a highly adaptive native perceptual system: Listeners can exhibit a small boundary shift to accommodate ambiguous tokens (e.g., Norris et al., 2003), or a bigger boundary shift in the face of a sound that falls unambiguously into an unintended category (e.g., Sumner, 2011), or structural adjustment within a category as shown here. More broadly, the adaptive ability in updating within-category structure may enable listeners to readily adapt to unfamiliar idiolects, dialects or even non-native accents in which speech tokens do not cause cross-category

confusion. Re-organization within the category proper may be sufficient to improve online speech processing in these scenarios, enabling not only disambiguation of tokens near the category boundary, but speeding lexical access to tokens nearer the category center. Future work is needed to test this possibility. For example, the Spanish vowel /u/ differs acoustically from English /u/ (Bradlow, 1995) but can be identified as the intended category easily by native-English listeners (e.g., Wade et al., 2007). When hearing the accented tokens, English listeners may update the best-exemplar region of the /u/ category during perceptual learning of Spanish-accented /u/ sounds, leaving the phonetic boundary location unchanged.

It is important to note that strong lexical-to-phonetic feedback is necessary for adaptation to occur. As shown in Experiment 7, when a speaker's productions were not very intelligible, each word provided less clear lexical information to help listeners interpret the speech signal. Although Kraljic and Samuel (2007) demonstrated that as few as ten critical words were sufficient to elicit phonetic recalibration, our results show that adapting to variable tokens in natural non-native accents is more effortful. Meanwhile, speakers with low intelligibility may also demonstrate larger deviations from the native distributions and require more drastic adjustments (e.g., a bigger shift in the boundary location). Future research should explore whether large deviations itself slow adaptation, with other factors being equal (i.e., when intra-talker variability and the amount of lexical-to-phonetic feedback are controlled).

At Which Level does Adaptation Occur?

The *second* major finding is that perceptual learning operates at a somewhat finer-grained level than "phonological abstraction" (Norris et al., 2003): undoubtedly listeners have abstracted away from the specific lexical items and generalize whatever they learn during exposure to novel words at test, but they also show sensitivity to the precise acoustic cues that are used in phoneme

specification. Listeners' attention to acoustic-level detail is manifested in several experiments. In Experiment 1, acoustic analysis showed that Speaker 1's productions of /d/ and /t/ did not entirely overlap; rather, the cue (burst length) he used to distinguish voicing was not an informative cue typically used by native listeners. Experiment 1 provided evidence that in the face of phonetic variation beyond native regularities, native listeners dynamically updated their own cue-weighting functions. Much of the discussion in perceptual learning studies has focused on the abstractness of pre-lexical retuning, given the evidence that learning generalizes across the lexicon (Norris et al., 2003; McQueen et al., 2006), across word positions (Jesse & McQueen, 2011; Eisner et al., 2013) and across places of articulation (Kraljic & Samuel, 2006). However, Reinisch et al. (2014) found that native perceptual learning, at least when guided by visual information (a visual /aba/ or a visual /ada/), was restricted to the specific acoustic cues to which listeners were exposed. If the specific cues used to contrast /b/ and /d/ were not aligned between exposure stimuli and test stimuli, for instance, /b/ and /d/ were contrasted by formant transitions in exposure and were then distinguished by burst information at test, then recalibration of the phonetic boundary did not generalize to test stimuli. In line with Reinisch et al. (2014), our data highlight an important role of specific acoustic cues in reorganizing the internal structure of phonetic categories as listeners adapted to a specific talker.

Results from Experiments 1 and 6 together further demonstrated that listeners were highly sensitive to the within-talker consistency of specific acoustic cues in cueing phoneme identity; high within-talker variability was observed to have negative impacts on rapid adaptation. This is a novel finding. In previous studies on perceptual adaptation, production variability is relatively neglected as a topic of investigation. Naturally, speaker intelligibility is often correlated with within-category variability in production (e.g., Wade et al., 2007). We

controlled for speaker intelligibility in Experiments 1 and 6 in order to understand the role of acoustic variability. The current finding is consistent with a number of findings on phoneme categorization in native speech. In an eye-tracking study, Clayards et al. (2008) demonstrated that listeners were less certain in categorizing stops (/b/ or /p/) along a VOT continuum and showed shallower categorization curves when within-category cue variability was large, even when the central tendency of acoustic distribution was held constant. This suggested that, to be sensitive to the *variance* of the distribution, listeners must have been tracking the *entire* probabilistic distribution of the phonetic category. In Clayards et al. (2008), the amount of within-category variability was coupled with the amount of between-category overlap. Other research has shown that within-category variability independently contributes to perceptual uncertainty and identification difficulty (Newman et al., 2001; Hazan et al., 2013). The current results extend this set of findings by showing that when the acoustic distributions of the speech signal do not match existing representations, listeners dynamically adapt by building up new cue-category mappings for a specific talker. Crucially, the remapping not only contains information about where the category boundary is (as shown in studies of phonetic boundary recalibration and Experiment 2), and about what are good tokens (as shown in Experiment 3) but also the amount of distributional variation for the talker. As noted in the discussion of Experiment 6, the data were supportive of listeners' tracking of talker-specific productions, rather than overall stimulus variability.

When do Listeners Generalize across Talkers?

Support for attention to acoustic detail in adaptation also comes from our finding on cross-talker generalization. In Experiment 4, talker generalization of phonetic retuning was governed by bottom-up similarity between talkers, instead of listeners' judgments of talker

identity or accent type. In addition, we found that even speakers of the same non-native accent demonstrated different subphonemic patterns (compare Speakers 1 and 2, who were matched on intelligibility), although at the phoneme level, their productions are somewhat similar (pronouncing word-final /d/ similar to /t/). Listeners tracked fine-grained detail along multiple dimensions in a rich acoustic-phonetic space, and they did not generalize experience with prior foreign-accented speaker to another speaker of the same accent if the speakers were not sufficiently similar along those acoustic dimensions. Thus, *acoustic similarity*, which contains richer information than an overall degree of segmental ambiguity or speaker intelligibility, constrains generalization.

Experiment 5 (5A and 5B) is the first study to investigate the benefits of multiple-talker exposure at pre-lexical levels. The results from the cross-modal priming task provide confirmatory evidence that an altered sound-to-category mapping underlies cross-talker generalization. The observed improvement in word recognition for a novel speaker was not due to a general relaxation that merely includes more competitors as a viable match to existing word forms. Meanwhile, our results show that even when listeners generalize prior experience to novel talkers, it does not necessarily mean that they have formed more *abstract*, talker-independent representations. First of all, listeners apparently did not develop an explicit awareness of a shared accent between accented speakers in the current work. Admittedly, the exposure phase was brief and listeners heard only isolated words. Compared to other intelligibility studies (Bradlow & Bent, 2008; Sidaras et al., 2009), listeners in Experiment 4 and 5 had fewer speech samples to evaluate the accent of speakers. It is possible that under more natural situations where listeners receive extraneous information about talkers' accents, they could develop an explicit knowledge

of the accent of speakers, and use it to actively predict incoming acoustic patterns and constrain generalization.

Second, our data indicated that listeners did not indiscriminately generalize to any novel talker of the same accent. This suggests that if some "implicit" knowledge of a shared accent boosts generalization across talkers, this implicit knowledge has to develop at the level of specific acoustic cues. Interestingly, in the Multi 1 → Speaker 1 condition where we observed talker generalization, Speaker 1 was not only acoustically similar to the exposure speakers in aggregate, but also highly resembled a particular speaker (Speaker 4). As discussed in Chapter 4, a "generalization-by-exemplar comparison" model is logically consistent with current finding. It is also consistent with all previous studies that show discrepant generalization patterns for fricatives and stops. These studies have examined transfer of phonetic retuning from one single speaker to a novel speaker. In every case where generalization was observed, talkers were indeed acoustically similar, if acoustic information was even reported (Kraljic & Samuel, 2007; Reinisch & Holt, 2014). It is noteworthy that in the literature of adult second language learning, *high variability training* approaches have been widely found to be most effective in helping adults acquire non-native phonetic categories and generalize to stimuli outside the training set (see Bradlow, 2008 for a review). However, in some cases, training with an appropriate individual talker was as effective as multiple-talker training in promoting stimulus-general learning and transferred to a novel talker (Magnuson et al., 1995). The notion of exemplar-type generalization is consistent with this kind of data.

Of course, it is possible that listeners may move from exemplar-type generalization to extraction-type generalization as they accumulate experience with acoustics of an accent. For instance, initial exposure to Mandarin-accented English may present information about /t/-like

/d/s or even specific constellation of vocalic cues and burst cues for word-final /d/. More and more exposure in the form of multiple talkers may draw listeners' attention to burst length as a phonetically-critical cue for the critical segment such that ultimately, it does not require an exact acoustic match along all acoustic dimensions to elicit generalization. Listeners may infer from the presence of a long burst that a novel speaker has a Mandarin accent and that they should apply their previous experience with Mandarin-accented speakers to understand this speaker. The extraction-type generalization may give listeners more flexibility and greater chance to reach stable speech perception in a dynamically-changing multi-linguistic society.

What are the Effects of Phonetic Reorganization on Lexical Access?

It is evident in the present study that the foreign-accented spoken words did not function fully as native words in the sense that a robust lexical activation was still observed for phonetically similar words. Following adaptation, accented /d/-final words (e.g., 'seed', pronounced like 'seat') activated the intended word form more strongly; however, phonetically-related competitors ('seat' for 'seed') were not eliminated from consideration. This pattern was observed for talker-specific adaptation (Experiment 1) and generalization to a novel talker (Experiment 5B). These results are somewhat in contrast with studies on adaptation to native dialectal variants (Dahan, Drucker, & Scarborough, 2008; Trude & Brown-Schmidt, 2012). Trude and Brown-Schmidt (2012) found that as listeners adapted to a dialectal speaker who pronounced the vowel of 'bag' as '[eɪ]' (as in 'bake'), they were more quickly eliminating 'bag' as a competitor for 'back' than they were accepting 'bag' as a potential candidate for the auditory 'bake'. Our results were the opposite: listeners included an accented /d/ as more acceptable candidates of a lexical entry (e.g., 'seed') before they eliminated the accented token as viable candidates for /t/-final words (e.g., 'seat'), if a complete elimination of phonological competitors

could ever be achieved. This may imply that listeners are more conservative facing a foreign-accented speaker whose productions mismatch native categories not just for a single sound in specific words, but mismatches native speech more globally for multiple consonants and vowels. Another potential reason why learning for a foreign accent seems incomplete (within the time frame examined) compared to learning an altered phoneme in a native accent (cf. McQueen et al., 2006) may lie in the specific process that achieves the adaptation. That is, shifting attention to a previously unattended acoustic dimension may require more perceptual and/or cognitive efforts than recalibrating a boundary along a familiar dimension does (e.g., recalibrating /s/ and /f/ boundary, or a VOT boundary between /d/ and /t/).

Theoretical Implications and Future Directions

A consistent finding across all three sets of experiments is that even in this brief exposure paradigm, listeners tracked acoustic distributions along multiple dimensions in a speaker's productions and adjusted the mapping from sounds to words accordingly, as guided by lexical information. This ability is crucial in adaptation to unfamiliar pronunciations that deviate from native norms as listeners used this information to a) rapidly restructure phonetic categories and facilitate word recognition for a given talker; b) effectively "assess" whether generalization to a novel talker is appropriate. Moreover, the degree of within-talker acoustic variability, not just the presence of it, predicted talker-specific adaptation results.

As discussed in Chapter 2, neither the talker normalization theories which assume invariant, canonical phonological representations for all talkers, nor the episodic theories which assume a detailed representation for each word by each talker, provide sufficient principles to account for existing evidence on perceptual adaptation studies. Several modified versions have been proposed to chart a middle course between the extreme stances (Mirman et al., 2006;

Goldinger, 2007; Pierrehumbert, 2006; Johnson, 2006). Of special interest here are the models that specify how talker adaptation and its generalization work. Mirman et al. (2006) added a Hebbian learning algorithm to the TRACE model (McClelland & Elman, 1986), allowing the abstractionist model to adjust the connections from acoustic features to phoneme units, as a way to adapt to speaker characteristics. The Hebbian-TRACE model demonstrated that bottom-up acoustic similarity (or lack of it) in phonetically-relevant cues constrained generalization. However, as the model has to adjust constantly as the talker changes, without a mechanism to store the learning results, it cannot account for persistent effects of talker adaptation (e.g., Eisner & McQueen, 2006; Kraljic & Samuel, 2005). Furthermore, because the model does not discriminate intra- versus inter-talker variability in the speech input, it would predict that both types of variability block, or at least slow down rapid adaptation. However, as we noted in the discussion of Experiment 6, the variability coming from Speaker HV's productions and that from multiple talkers did not produce the same effect, implying that listeners were sensitive to the source of variability.

Johnson (2006) proposed an exemplar-resonance model that allows explicit recognition of social identity of talkers (e.g., gender) and uses the social label to bias speech recognition. In this model, listeners retain speech exemplars in memory that contain information about both linguistic categories (words) and social categories (e.g., female and male). The incoming speech signal activates existing exemplars that are acoustically similar, and the activated exemplars further activate categorical information (a word "seed" or a male speaker). In addition, once an exemplar is activated, feedback activation from categories to exemplars (resonance) spreads activation out to other exemplars that share the category membership. This model would readily account for a few findings. First, since acoustic similarity determines the overall exemplar

activation for a given category, it would predict that generalization occurs only when bottom-up similarity supports it. It would also predict that the more exemplars of accented "seed" that are similar to the input, the larger activation for "seed". This is consistent with the *generalization-by-exemplar* hypothesis we suggested in explaining the difference between the Multi 1 → Speaker 1 and Speaker 2 → Speaker 1 conditions. Second, since talker detail is stored in exemplars, listeners naturally discriminate between variability within- and across-talkers, which were found to constrain generalization differently. However, a problem remains for this episodic model: if no abstract pre-lexical representation of phonetic categories is used in spoken word recognition, how could listeners possibly generalize from a few limited speech samples to novel words? Considering the fact that a Mandarin-accented "seed" would potentially activate pre-existing English-accented "seat" exemplars (which should outnumber Mandarin-accented "seed" exemplars due to long-term native language experience), it is especially hard for this model to account for the rapid adaptation and generalization.

The current finding is consistent with a framework in which listeners a) routinely track the full distribution (not only the mean, but also the variance) of acoustic-phonetic cues across individual talkers, b) dynamically weight each cue accordingly to its informativeness to phoneme distinction, for a given talker and c) use the similarity in acoustic-phonetic cues to constrain talker generalization for an adjusted cue-to-category mapping. Two types of models are suitable to account for the data.

The Attention-to-Dimension (A2D) model (Francis et al., 2000; Francis & Nusbaum, 2002) for non-native language learning can be adapted to account for the rapid adaptation data. This model acknowledges that non-native languages differ from native language not only in the absolute distribution along familiar dimensions (e.g., VOT differences for French stops and

English stops), but also in the particular acoustic dimensions that reliably cue phoneme contrast. For instance, Japanese listeners have difficulty distinguishing English /r/ and /l/ because they rely on both the second and third formant frequency (primarily F2) as cues for distinction, whereas the phonemes are predominantly contrasted by F3 in English. A mismatch in the cue weighting strategy between an English speaker and a Japanese listener causes perceptual difficulty (e.g., Yamada & Tohkura, 1992). Given category labels as feedback, listeners can be trained to selectively weigh particular cues more heavily in phoneme identification (Francis et al., 2000) or to attend to previously unattended cues (from VOT to F0 in distinguishing Korean stops, Francis & Nusbaum, 2002). It is made explicit in the A2D model that selective attention to relevant cues in the L2, which might be less informative in one's L1, is helpful in acquiring new phonetic categories and generalizing to new syllabic contexts. A similar discrepancy of cue-weighting functions arises between a foreign-accented speaker and a native listener. In the present study, the Mandarin-accented speaker (Speaker 1) did not saliently vary vowel duration in the voicing contrast. In adaptation, English listeners who were exposed to a few /d/-final words showed evidence of relying more on the burst length than the control group, a cue that is not typically used to make the voicing distinction in English. It is important to note that the current experiments did not provide a rigorous within-subjects test to see if sensitivity is indeed increased for certain acoustic cues (pre- and post-adaptation) as examined in studies of cue-weighting in learning non-native categories (e.g., Francis & Nusbaum, 2002). Future research is required to directly address the role of attention in perceptual adaptation. Although native phonetic retuning seems effortless, evidence exists that phonetic retuning may be disturbed when cognitive load is synchronously high (Samuel, 2014). Meanwhile, increased processing effort is usually required for foreign-accented speech (e.g., Schmid & Yeni-Komshian, 1999; Munro &

Derwing, 1995). It remains unclear how crucial it is to effectively allocate attention to the relevant aspects of speech input in accent adaptation.

While the A2D model is an appealing model to account for the current findings, the model itself does not readily generate predictions about the time course of different types of perceptual learning. Apparently, rapid phonetic adaptation observed here occurred over much shorter time than shown in previous training studies of non-native phoneme learning. It is possible that the amount of deviation from the native language (which is likely larger in a non-native language than in a foreign accent than in an idiosyncratic native variant, e.g., Kraljic & Samuel, 2007) is the determining factor for the time course of learning. Additionally, shifting attention between acoustic dimensions might require more time than adjusting the boundary location along a pre-attended dimension.

Another issue concerns the mechanism of generalization. In past research, learning-elicited selective attention to acoustic cues has only been demonstrated in talker-specific (e.g., Francis & Nusbaum, 2002) and context-specific conditions (e.g., visually-guided phonetic adaptation, Reinisch et al., 2014). In the current study, similarity along specific dimensions between talkers predicted generalization. However, our design could not assess whether listeners applied the adjusted weighting to the novel talker. Related to this, Witteman et al. (2013) found that exposure to a German-accented Dutch speaker did not immediately facilitate word recognition of a second accented speaker, but somehow made the adaptation for the second speaker faster (compared to listeners who did not have experience with the first speaker). It could be that in the face of a different speaker, listeners did not directly apply the adjusted representation (due to perceived acoustic dissimilarity in the distributions of a cue, for example). Yet exposure to the first speaker could have retuned listeners' attention to relevant acoustic

information that would be helpful in the recognition of the critical phoneme, and thus made the adaptation to the second speaker faster. The current study could be easily modified to provide a test of the generalization issue. First, we could adopt an orthogonal design (Holt & Lotto, 2006) such that for the exposure stimuli, *vowel* length (deviating from native values) serves as the only informative cue for one group of listeners while *burst* length serves as the only informative cue for another group. We then test each group on two sets of stimuli, containing the cue *consistent*, or *inconsistent* with the exposure stimuli. All stimuli would be created out of Speaker 1's natural productions. The A2D model would predict that generalization from exposure to test would be observed only in the *consistent cue* condition. That is, even for a given speaker, generalization is not guaranteed if different cue-weighting changes are required for specific stimuli. Moreover, since vowel length is a cue that English listeners typically use, adaptation for the *consistent vowel cue* condition is predicted to be faster than the *consistent burst cue* condition. Complementary to these predictions, we predict that the re-weighting of acoustic cues, in particular the added weighting of burst length for the Mandarin accent, would allow listeners to generalize their learning to other stop consonant contrasts (e.g., /b/-/p/). Generalization from one accented speaker to another would also be observed if they use the same set of acoustic cues. Future research should investigate whether the observed adjustment in the phonetic structure and cue-weighting strategy generalize to other contexts (e.g., phonemes, speakers, accents).

As noted above, attentional shifts between acoustic cues may help listeners to move from exemplar-type generalization to extraction-type generalization as they gain more experience across multiple talkers which allow them to pay attention to the most talker-general, relevant cues (and ignore irrelevant acoustic variation) in an accent. At its heart, this account aligns with the "distilling regularities" idea that was hypothesized by Bradlow and Bent (2008) and Sidaras

et al. (2009). We did not find sufficient support for it within our study. Future work should test whether more extensive exposure might promote it.

Bayesian models can be used to model the current findings. Clayards et al. (2008) tested an *ideal observer* model which makes use of "entire probability distributions" and also involves cue weighting in speech perception. It explicitly predicts a role of acoustic variance (on top of mean values). In Bayesian terms, the task of speech perception is to determine the probability of a category identity, given an input cue (posterior probability). For instance, observing a burst length of 40 ms, what is the probability of category /d/? This posterior probability is proportional to the prior probability of a category (e.g., the likelihood of /d/) times the conditional probability of a cue input given the category (i.e., the likelihood of a 40ms burst given a production of /d/). The outcome of phoneme identification depends on the posterior probability of /d/ (relative to posterior probability of other categories, e.g., /t/). In the current study, the word frequency of /d/-final words and /t/-final words were equated. That is, prior probabilities of /d/ and /t/ were the same. As listeners encounter a Mandarin-accented speaker, they build up the conditional probability in his productions, based on the exposure stimuli. A widely distributed acoustic-phonetic cue decreases the likelihood each value of the cue (e.g., a bust length of 40ms) appears as a member of a category (e.g., /d/). That is, the conditional probability gets smaller. And thus, posterior probability of /d/ gets smaller. This explains the data that a speaker with large within-talker variability was harder to adapt to.

In an advanced version of the model, Kleinschmidt & Jaeger (2014) proposed an *ideal adapter belief updating* framework that builds on the same Bayesian concept but allows updating of conditional probabilities that is sensitive to talker information. The technical details are beyond the scope of this dissertation, but two assumptions of the framework should be noted.

First, listeners build up *generative* models of category-to-cue mappings via statistical learning. For instance, they know how well a /d/ predicts a burst of 40ms. This is shown in Clayards et al. (2008). This level of Bayesian learning explains how talker-specific adaptation occurs. Second, the generative models are tailored to specific talkers. Listeners build up separate talker-specific generative models based on their prior experience with that talker. That is, talker is a parameter in computing the posterior probability. In addition, listeners also estimate how likely one or more previously-built generative model matches the productions of a new talker. In a sense, talker identity is represented as distributions of generative models. At this level, instead of computing probabilities of categories given cue input, listeners compute the probability of a generative model given an observed input and a talker (whose identity is known). This level of Bayesian learning allows predictions of talker generalization. Pertinent to the current finding, this framework predicts that listeners use previous generative models to narrow down the interpretation of tokens from a novel talker. The more overlap between the acoustic distributions of the current speech input and a previous generative model, the more likely a generative model will be used to interpret the current speech input. It would be harder to adapt to a novel talker whose generative model falls out of the range of previous models. Our finding that listeners generalized in the Multi 1 → Speaker 1 condition but not Multi 2 → Speaker 2 condition is consistent with this prediction. Needless to say, the model also predicts that there is no generalization from one talker to another if they are not acoustically similar (that is, a previous generative model will not be applied to the productions of a novel talker), which is exactly the case in the Speaker 1 → Speaker 2 and Speaker 2 → Speaker 1 condition. Lastly, because the inference of generative models is uncertain for any talker, generalization does not require explicit knowledge of talker or accent identity, as found in the current study. The framework additionally

predicts that when such social information is available, listeners will narrow down the selection of generative models more easily. Trude et al. (2013) showed that providing contextual cues to talker identity helped listeners constrain adaptation to a specific talker in a two-talker situation. No direct evidence is available for the multiple-talker situation. Future studies should investigate whether "telling" listeners (e.g., providing long carrier sentences or visual information, etc.) that they are listening to a group of Mandarin-accented speakers would facilitate adaptation at a group level and promote generalization to a novel talker, keeping bottom-up input consonant. A strength of this framework is that by constructing different acoustic distributions, the computational model can simulate outcomes that are testable in human participants. One important issue suitable for computational testing is how the exact type of phonetic adjustment (boundary shift, within-category reorganization, or attentional shifts between acoustic dimensions) depends on the distributional properties of the acoustic variation and the time course of each type of adaptation.

Taken together, the findings presented in this dissertation practically bridge two lines of research which have been conducted separately: the lexically-guided phonetic reorganization mechanism that underlies the adaptation to idiolect differences of native speakers also supports adaptation to natural foreign accents. In addition, bottom-up similarity at the acoustic-phonetic level explains a range of situations in which adaptation effects may or may not generalize to novel talkers, reconciling some of the inconsistent results in both domains. We expand previous research by showing that perceptual learning for speech is not just a matter of adjusting phonetic boundaries in face of noncanonical tokens; it also promotes a reorganization of the internal category structure. Furthermore, the ability to track acoustic-phonetic characteristics for individual talkers helps listeners dynamically adapt and selectively generalize, maintaining a

balance between flexibility and stability in a highly variable acoustic environment. I believe the future directions suggested above would further advance our understanding of the reorganization of the perceptual architecture that listeners experience when they adjust to accented speech, and other types of unfamiliar speech in general.

REFERENCES

Agresti, A. (2002). *Categorical data analysis*. New York: John Wiley & Sons.

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, *63*, 798-810.

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

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*, 163-187.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*, 255-278.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*, 390-412.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database [CD-ROM]. Philadelphia: Linguistic Data Consortium.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America*, *133*, EL174-EL180.

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica*, *65*, 131-147.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech*

learning: In honor of James Emil Flege, 13-34.

Bradlow, A. R. (1995). A comparative acoustic study of English and Spanish vowels. *The Journal of the Acoustical Society of America, 97*, 1916-1924.

Bradlow, A. R. (2008). Training non-native language sound patterns: lessons from training Japanese adults on the English. *Phonology and Second Language Acquisition, 36*, 287-308.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*, 707-729.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & psychophysics*, *61*(2), 206-219.

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539-571.

Clarke, C. M. (2000). Perceptual adjustment to foreign-accented English. *Journal of the Acoustical Society of America, 107*, 2856 (A).

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.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804-809.

Connine, C. M., Blasko, D. G., & Wang, J. (1994). Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context. *Perception & Psychophysics, 56*, 624-636.

Crowther, C.S., & Mann, V. (1992). Native language factors affecting use of vocalic cues to final consonant voicing in English. *Journal of the Acoustical Society of America, 92,* 711-722.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America, 83,* 1553-1573.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations?.*Cognition, 108,* 710-718.

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

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

Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology, 4,*148.

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

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

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America, 106*(5), 2973-2987.

Flege, J., Munro, M., & Skelton, L. (1992). Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin and Spanish, *J. Acoust. Soc. Am., 92,* 128-143.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing?. *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1276-1293.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*, 1668-1680.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 349.

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

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive psychology*, *45*, 220-266.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, *34*, 65-87.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, *22*, 1166.

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

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th international congress of phonetic sciences* (pp. 49-54).

Hazan, V., Romeo, R., & Pettinato, M. (2013, June). The impact of variation in phoneme category structure on consonant intelligibility. In *Proceedings of Meetings on Acoustics* (Vol. 19, No. 1, p. 060103). Acoustical Society of America.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059-3071.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic bulletin & review*, *18*(5), 943-950.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perception. *Language and speech*, *34*(3), 265-279.

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

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of phonetics*, *34*(4), 485-499.

Jongman, A., Wade, T., & Sereno, J. (2003). On improving the perception of foreign-accented speech. *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1561–1564). Barcelona, Spain.

Kakehi, K. (1992). Adaptability to differences between talkers in Japanese monosyllabic perception. *Speech Perception, Speech Production, and Linguistic Structure*, 135-142.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208-1221.

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.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54-81.

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

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

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

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts how listeners adjust to speaker variability. *Psychological science*, *19*(4), 332-338.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics*, *50*(2), 93-107.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, *51*(6B), 2018-2024.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning

new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242-1255.

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

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, *78*(6), 1949-1957.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), Cognitive models of speech processing: Psycholinguistic and computational perspectives (pp. 122-147). Cambridge, MA: MIT Press.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391.

Magnuson, J. S., Yamada, R. A., Tohkura, Y., & Bradlow, A. R. 1995. Testing the importance of talker variability in non-native speech contrast training. *Journal of the Acoustical Society of America*, 97(5), Pt. 2: 3417.

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

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

Marslen-Wilson, W., Nix, A., & Gaskell, G. (1995). Phonological variation in lexical access: Abstractness, inference and English place assimilation. *Language and Cognitive*

Processes, 10, 285–308.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological review, 101*(4), 653.

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.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86.

McGowan, R. S., & Cushing, S. (1999). Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis. *The Journal of the Acoustical Society of America, 106*(2), 1090-1105.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science, 30*(6), 1113-1126.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America, 131*(1), 509-517.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*(6), 457-465.

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.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic bulletin & review*,*13*(6), 958-965.

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.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088-2113.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181-1196.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, *115*(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(03), 299-325.

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

Nusbaum, H. C., & Henly, A. S. (1992). Listening to speech through an adaptive window of analysis. *The Auditory Processing of Speech: From Sounds to Words*, 339.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. *Talker variability in speech processing*, 109-132.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. *Speech perception, production and linguistic structure*, 113-134.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42-46.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory.*Perception & Psychophysics*, *57*(7), 989-1001.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175-184.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *34*(4), 516-530.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories.*Attention, Perception, & Psychophysics*, *15*(2), 285-290.

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.

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

Rochet, B. L., & Yanmei, F. (1991). Effect of consonant and vowel context on Mandarin Chinese VOT: production and perception. *Canadian Acoustics*, *19*(4), 105-106.

Rogers, C.L. (1997). Intelligibility of Chinese-accented English (Doctoral dissertation, Indiana University, 1997). *Dissertation Abstracts International*, *58*, 9.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, *31*(4), 307-314.

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, 6249-72.

Samuel, A. G., & Kraljic, T. (2013). *Visually specified speaker identity can dominate processing of spoken words*. Manuscript submitted for publication.

Samuel, A. G. (2014). How Much Processing Time Is Needed to Drive Perceptual Recalibration of Speech?. Annual meeting of the Psychonomic Society, Long beach, CA. November,

2014.

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research*, *42*(1), 56-64.

Sidaras, S. K., Alexander, J. E. D., and Nygaard, L. C. (2009). "Perceptual learning of systematic variation in Spanish accented speech," *J. Acoust. Soc. Am. 125*(5), 3306–3316.

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

Skoruppa, K. and S. Peperkamp (2011). "Adaptation to novel accents: feature-based learning of context-sensitive phonological regularities. *Cognitive Science 35*, 348-366.

Strand, E. A., & Johnson, K. (1996, October). Gradient and Visual Speaker Normalization in the Perception of Fricatives. In *KONVENS* (pp. 14-26).

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131-136.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974-3982.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *128*(4), 2090-2099.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434-464.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online

speech perception. *Language and Cognitive Processes, 27*(7-8), 979-1001.

Trude, A. M., Tremblay, A., & Brown-Schmidt, S. (2013). Limitations on adaptation to foreign accents. *Journal of Memory and Language, 69*(3), 349-367.

Utman, J. A., Blumstein, S. E., & Burton, M. W. (2000). Effects of subphonetic and syllable structure variation on word recognition. *Perception & Psychophysics, 62*(6), 1297-1311.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America, 92*(2), 723-735.

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica, 64*(2-3), 122-144.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of phonetics, 32*(2), 251-276.

Warren, P., & Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics, 41*(3), 262-275.

Warren, P., & Marslen-Wilson, W. (1988). Cues to lexical choice: Discriminating place and voice.*Perception & Psychophysics, 43*(1), 21-30.

Witteman, M. J., Bardhan, N. P., Weber, A., & McQueen, J. M. (2014). Automaticity and Stability of Adaptation to a Foreign-Accented Speaker. *Language and Speech, 58*(2), 168-189.

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.

Wong, P. C., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1173-1184.

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics*, *41*(5), 369-378.

Yamada, R. A., & Tohkura, Y. I. (1992). The effects of experimental variables on the perception of American English/r/and/l/by Japanese listeners. *Perception & psychophysics*, *52*(4), 376-392.

Zhang, X., & Samuel, A. G. (2013). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 200-217.

APPENDICES

Appendix A Intelligibility Tests for Mandarin Speakers

In a pilot study, all Mandarin speakers recorded a word list with 190 words (Rogers, 1997), each of which included one or more phonemes predicted to cause difficulty for native listeners when spoken with a Mandarin accent. The words sampled across vowels and consonants that are difficult for Mandarin speakers of English. A separate group of 24 listeners transcribed words from the 190-word lists of the six speakers to establish baseline intelligibility for each of the Mandarin speakers. Six counterbalanced lists were created. Results are presented in Table A1.

Critical Exposure /d/-final Words

The same 24 native English-speaking listeners also completed a 2AFC identification task to assess the ambiguity of the final consonant in the training -/d/ words of each Mandarin speaker. During this task, listeners were asked to decide whether the word they heard ended in /d/ or /t/. For example, for the auditory item *apprehend,* they were asked to choose between *apprehend* or *apprehent.* Likewise, for *apprehent,* they chose between *apprehend* and *apprehent.* It was emphasized to the listeners that they would hear both words and nonwords, and their decision should be based on the final sound only. Speaker 1 and speaker 2 were matched on their overall intelligibility and intelligibility of critical /d/-final words in the experiments. In addition, words from each speaker in the multiple-talker condition (Experiment 4 and 5) were selected in a way that equated the overall ambiguity of exposure words (% /d/ responses given for /d/-final words in the 2AFC task) across experiments.

Table A1. Average intelligibility scores (expressed in % words correctly transcribed), and performance for exposure /d/-final words for all Mandarin speakers. Numbers in parentheses are standard errors. Demographic information is represented in the last 3 columns.

Speaker	Overall intelligibility	/d/ responses (%) for exposure words	Age of English Acquisition (years)	Age of arrival in the U.S. (years)	Length of Residence (months)
1	46(1)	72(5)	11	18	18
2	46(2)	77(4)	12	26	42
3	37(1)	49(6)	10	15	36
4	34(2)	78(3)	11	19	24
5	70(2)	88(4)	7	24	60
6	64(1)	83(4)	12	22	6

Critical /d/-final Test Words from Speakers 1, 2 and 3

To establish ambiguity of /d/ pronunciations in critical test items, three separate groups of 10 native-English listeners completed a 2AFC identification task. Participant group 1, 2, 3 heard tokens from speaker 1, speaker 2 and speaker 3, respectively. In the task, participants heard a mixed list containing all 60 critical minimal pairs of /d/-final and /t/-final words. Two counterbalanced lists were created with each list blocked into two halves, such that either -/d/ (e.g., *seed*) or -/t/ word (e.g., *seat*) of each minimal pair was presented in each half. Within each half, items were repeated in three sub-blocks. Within each sub-block, items were randomly presented. Performance did not differ across subblocks or halves of the lists. Thus we report the mean percent correct response for /d/-final words as the intelligibility measure for each speaker. The mean intelligibility was comparable between speaker 1 (M = .65, SD = .13) and speaker 2 (M = .70, SD = .06), with no significant difference, $t(18) = 1.061$, p = .30. Speaker 3 (M = .22,

SD = .17) had lower intelligibility for /d/ tokens than Speaker 1($t(18) = 6.739$, $p < .001$) and 2 ($t(18) = 9.049$, $p < .001$).

Appendix B Experiment Materials

Table B1. Example stimuli in the cross-modal priming task in Experiment 1, 4-7. Auditory primes are in lower case; visual targets are in capital letters. There were four counterbalanced lists. Each list contained 15 groups of stimuli as shown in the table.

		List 1	List 2	List 3	List 4
Critical trials	Related prime /d/	seed-SEED	need-NEED	pod-POD	herd-HERD
	Unrelated prime /d/	smile-POD	pearl-HERD	house-SEED	fair-NEED
	Related prime /t/	herd-HURT	seed-SEAT	need-NEAT	pod-POT
	Unrelated prime /t/	fair-NEAT	smile-POT	pearl-HURT	house-SEAT
Filler trials	Related priming	foam-FOAM			
	Unrelated priming	male-HORN			
	Nonword target	ring-WELF	smash-ZOG		wing-ZID
		shawl-WEM	mom-YICK		hill-TOVE

Appendix C Experiment Results

Table C1. Response accuracy in the auditory lexical decision task (Exposure phase) across experiments. Critical words are /d/-final words for the experimental group and replacement words for the control group. Standard deviations are presented in parentheses.

Experiment	Exposure group	Critical words	Filler words	Nonwords
Experiment 1	Experimental	.81 (.09)	.80 (.06)	.77 (.16)
	Control	.81 (.09)	.84 (.05)	.69 (.17)
Experiment 4	Experimental	.83 (.07)	.82 (.07)	.71 (.15)
	Control	.75 (.10)	.81 (.08)	.68 (.15)
Experiment 5A	Experimental	.78 (.08)	.87 (.05)	.71 (.12)
	Control	.84 (.07)	.89 (.03)	.70 (.11)
Experiment 5B	Experimental	.79 (.09)	.87 (.06)	.67 (.17)
	Control	.82 (.07)	.88 (.05)	.70 (.12)
Experiment 6	Experimental	.82 (.15)	.78 (.14)	.66 (.20)
	Control	.78 (.10)	.80 (.08)	.68 (.12)
Experiment 7	Experimental	.61 (.10)	.85 (.07)	.76 (.12)
	Control	.81 (.06)	.87 (.05)	.80 (.09)

Table C2. Mean error rates and RT across participants in the cross-modal priming task as a function of exposure group in Experiment 1 in Chapter 3. Standard deviations are given in parentheses.

	/d/-final		/t/-final	
Exposure group	Related prime	Unrelated prime	Related prime	Unrelated prime
Example	seed-SEED	fair-SEED	seed-SEAT	fair-SEAT
Mean % error				
Experimental	10 (7)	15 (13)	7 (7)	12 (10)
Control	9 (8)	18 (11)	6 (6)	13 (7)
Mean RT (ms)				
Experimental	598 (89)	672 (115)	593 (96)	630 (100)
Control	577 (89)	616 (92)	553 (87)	588 (81)

Table C3. Mean error rates and RT across participants in the cross-modal priming task as a function of exposure group across experiments in Chapter 4. Standard deviations are given in parentheses.

		/d/-final		/t/-final	
	Exposure group Example	Related prime *seed-SEED*	Unrelated prime *fair-SEED*	Related prime *seed-SEAT*	Unrelated prime *fair-SEAT*
Exp.4			Mean % error		
	Experimental	12 (8)	13 (8)	9 (6)	8 (6)
	Control	13 (9)	10 (9)	8 (7)	8 (5)
			Mean RT (ms)		
	Experimental	591 (78)	632 (91)	564 (68)	592 (70)
	Control	604 (79)	642 (76)	590 (72)	621 (72)
Exp.5A			Mean % error		
	Experimental	10 (9)	17 (10)	6 (5)	10 (7)
	Control	12 (9)	17 (12)	8 (5)	11 (10)
			Mean RT (ms)		
	Experimental	577 (50)	626 (63)	571 (64)	593(62)
	Control	576 (52)	610 (60)	570 (49)	592 (43)
Exp.5B			Mean % error		
	Experimental	10 (9)	16 (10)	6 (5)	11 (7)
	Control	10 (9)	18 (12)	7 (5)	10 (10)
			Mean RT (ms)		
	Experimental	578 (59)	625 (75)	565 (71)	584 (46)
	Control	633 (80)	647 (89)	603 (79)	633 (66)

Table C4. Mean error rates and RT across participants in the cross-modal priming task as a function of exposure group in Experiment 6 and 7 in Chapter 5. Standard deviations are given in parentheses.

		/d/-final		/t/-final	
	Exposure group	Related prime	Unrelated prime	Related prime	Unrelated prime
	Example	*seed-SEED*	*fair-SEED*	*seed-SEAT*	*fair-SEAT*
Speaker HV		Mean % error			
	Experimental	10 (8)	12 (8)	6 (5)	8 (5)
	Control	11 (8)	14 (11)	8 (8)	11 (10)
		Mean RT (ms)			
	Experimental	610 (86)	648 (68)	589 (77)	611 (65)
	Control	595 (56)	617 (63)	585 (65)	607 (72)
Speaker LI		Mean % error			
	Experimental	15 (13)	16 (8)	7 (6)	6 (6)
	Control	10 (9)	15 (8)	6 (6)	9 (9)
		Mean RT (ms)			
	Experimental	636 (75)	645 (64)	576 (56)	621 (56)
	Control	614 (107)	625 (87)	540 (71)	593 (87)