Distribution Agreement

Signature:

______________________________ ____________________
Siyu Lin Date

The role of previous experience in the perceptual adaptation of accented speech

By

Siyu Lin

Master of Arts

Psychology

Lynne C. Nygaard, Ph.D.
Advisor

Daniel D Dilks, Ph.D.
Committee Member

Phillip Wolff, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola
Dean of the James T. Laney School of Graduate Studies

Date

The Role of Previous Experience in Perceptual Adaptation of Accented Speech

By

Siyu Lin

B.A., University of Tennessee, 2018

M.S., Villanova University, 2021

Advisor: Lynne C. Nygaard, Ph.D.

An abstract of a thesis submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Master of Arts in Psychology

2023

Abstract

The Role of Previous Experience in Perceptual Adaptation of Accented Speech

By Siyu Lin

Adaptation to accented speech has been a long-standing problem in the field of speech perception, but the underlying mechanisms remain unclear. In recent years, accounts of adaptation to accented speech posit an exemplar-based mechanism, in which the extent to which listeners successfully generalize learning of accented productions to other accented talkers depends on the acoustic similarity between the two talkers. However, listeners' prior perceptual experience may also be involved in the process of adaptation. This study investigated whether listeners' prior perceptual experience would facilitate cross-talker generalization. Experiment 1 was designed to replicate and extend previous findings showing perceptual learning of /d/-final productions in Mandarin-accented English (Xie et al, 2017). Listeners were asked to perform a lexical decision task in an exposure phase and then tested in a cross-modal priming task with auditory stimuli produced by the same talker. Experiment 2 was designed to determine if the linguistic experience would influence the extent of learning. Two experiments (2a and 2b) were run in which the critical *exposure* words had vowel durations that were artificially extended by 80 ms (2a), and the other had critical *test* words with vowel durations that were artificially extended by 80 ms (2b). We expected that listeners would be more likely to perceptually adapt and generalize to productions that changed from less to more English-like productions between the exposure and test. Overall, we failed to replicate the perceptual adaptation effect reported previously, potentially because the acoustic difference between the critical exposure words and test words was larger than in previous work. We generated updated predictions for Experiment 2 to examine the degree of acoustic difference across exposure and test. Even though the results

again did not provide evidence for learning, numerical trends were consistent with a role for acoustic similarity in the generalization of learning.

The Role of Previous Experience in Perceptual Adaptation of Accented Speech

By

Siyu Lin

B.A., University of Tennessee, 2018

M.S., Villanova University, 2021

Advisor: Lynne C. Nygaard, Ph.D.

An abstract of a thesis submitted to the Faculty of the James T. Laney School of
Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of
Master of Arts in Psychology

2023

Table of Contents

Introduction

Speech is a highly variable signal. Listeners have to overcome all types of variation to extract the meaning correctly. When people listen to speech, there are two types of information that they can gain, linguistic information (e.g., sentences, words, and phonemes) and indexical information which pertains to the characteristics of the talker's voice. Indexical properties provide information about 1) group membership (Borrie, McAuliffe, & Liss, 2012; Bradlow & Bent, 2008; Clopper & Pisoni, 2007), 2) individual characteristics such as size and shape of vocal tract, which influences spectral properties of the speech signal such as formant frequencies (Johnson, Strand & D'Imperio, 1999; Ladefoged & Broadbent, 1957; Maye, Aslin, & Tanenhaus, 2008) and characteristic speaking rate, which influences temporal properties such as voice onset time (VOT: Theodore, Miller, & DeSteno, 2009), and 3) changing states of a talker such as emotional or attitudinal state (Scherer, 2003). All these indexical properties introduce variation in the way in which speech is produced and ensure that speech is uniquely produced by each individual talker. A longstanding problem in the study of speech perception is understanding how listeners contend with these sources of variation in spoken language to reliably recover linguistic structure.

Variations at talker- and group-specific level

Individual talker characteristics constitute one of the main sources of variation in speech. Early work provided evidence that talker-specific perceptual tuning could occur at the level of phoneme categories. Ladefoged and Broadbent (1957) created carrier phrases in which the final word (i.e., /b?t/) contained one of four vowels (i.e., /i/, /e/, /a/ and /ə/) and subjects were asked to identify these words. Critically, the carrier phrases were presented in two different synthesized voices and the results showed that the subjects identified the vowel differently depending on which voice they heard. The study suggests that listeners change their expectations and perceptual spaces based on the particular acoustic-phonetic characteristics of the talkers' voices that they encounter. Nygaard, Sommers and Pisoni (1994) exposed subjects to ten talkers' voices over the course of nine days of training. Then, at test, subjects were asked to identify spoken words produced by either familiar or novel talkers. The results showed that subjects were better at identifying words produced by the familiar talkers, which suggests that experience with the specific set of talkers' voices helped to tune the subjects' perceptual mechanisms to better recover the linguistic structure in the speech signal.

Talkers can share characteristics such as accent or dialect and there is evidence that listeners adapt to group-based characteristics as well. For example, speakers of the same dialect or accent group share characteristic pronunciations that both identify them as members of that group as well as introduce variation in the way speech sounds are realized (Clopper & Pisoni, 2004, 2007). More importantly, the group-based variations can be adapted to by listeners. Clarke and Garrett (2004) exposed native speakers of American English (AE) to spoken sentences and asked whether a visual probe presented on screen matched the sentence-ending word. Listeners heard the sentences produced in either a Mandarin accent (accent condition) or in American English (AE during exposure but heard only Mandarin-accented sentences (produced by the same talker as in the Mandarin exposure condition) at test. The results showed that the listeners who were exposed to the Mandarin talker responded faster and were more accurate relative to listeners exposed to American English speech. Given that the duration of the experiment was as short as a few minutes, the study suggests that listeners are able to quickly adapt to accented speech. In a replication by Xie et al. (2018), the results showed that successful generalization could occur even though the test stimuli were produced by a different Chinese-accented talker.

Although an accent can be seen as one dimension of a talker's individual indexical characteristics, it also implies group membership because its acoustics varies systematically across a group of talkers. Therefore, it is interesting to investigate whether listeners are able to create a *talker-independent* representation of a particular accent. Bradlow and Bent (2008) exposed subjects to Chinese-accented sentences for two sessions over the course of two days in a sentence transcription task and then asked them to transcribe sentences produced by either a new Chinese-accented or Slovakian-accented talker at test. The subjects showed better transcription accuracy at test when the test talker's accent was Chinese than when it was Slovakian. Also, subjects who had multi-talker exposure performed better than the ones who had only the exposure to a single talker. Together, these findings suggest that listeners could create representations for a specific accent (i.e., Chinese accent) and that multi-talker exposure could help to facilitate the perceptual learning process. Sidaras, Alexander and Nygaard (2009) provided evidence that listeners who were exposed to Spanish-accented English produced by multiple talkers generalized learning to novel words and sentences produced by unfamiliar Spanish-accented talkers, which enhances the hypothesis that listeners are able to obtain a talker-independent and accent-specific representations (see also Alexander & Nygaard, 2019). However, the downside of most of these studies is that they used intelligibility of the auditory stimuli as a global measure of adaptation to the accented speech. Therefore, the perceptual mechanisms involved in adaptation remain unclear.

Underlying mechanisms of speech adaptation

Phonetic retuning in which listeners tune their phonetic category structure to particular types of variation may be one underlying mechanism of perceptual adaptation in spoken language (Norris, McQueen & Cutler, 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006, 2007, 2009; Melguy & Johnson 2022; but see Zheng & Samuel, 2020). Norris, McQueen and Cutler (2003) exposed subjects to ambiguous /s/-/f/ sounds in spoken words that were consistent with one or the other meaningful interpretation of the fricative (e.g., /s/-biasing *bliss*; /f/-biasing *surf*). Ambiguous fricatives were sampled at the middle point on the /s/-/f/ continuum (e.g., the fricative is manipulated such that its acoustic value is equally different from either /s/ and /f/). These items were presented along with filler words and nonwords and subjects were asked to judge whether each item that they heard was a word or not. Following exposure, subjects were asked to categorize fricatives drawn from acoustically equivalent steps on an /s/-/f/ continuum. The results showed that the categorization boundary shifted depending on lexical guidance such that the group that heard the ambiguous fricative in /s/-biasing contexts categorized more steps as /s/, and the group that heard the ambiguous fricative in /f/-biasing contexts categorized more steps along the continuum as /f/. Further evidence suggests that the boundary shift constitutes a re-tuning of the entire perceptual category structure (Clarke-Davidson, Luce & Sawusch, 2008; Xie, Theodore & Myers, 2017) during adaptation to the ambiguous or atypical speech sound.

Eisner and McQueen (2005) used the lexically guided perceptual learning paradigm to examine whether this kind of phonetic retuning exhibits talker specificity, such that learning is specific to an individual talker’s pronunciations. In their study, the boundary shift indicating perceptual adaptation was observed when the fricative continuum at test was based on the exposure talker but not when the continuum at test was produced by a different talker. The results showed that speech adaptation was talker-specific and also suggested that talker-specificity can be represented at the critical segment (i.e., fricative; see also Kraljic & Samuel, 2007; Kraljic, Samuel & Brennan, 2008). Reinisch and Holt (2014) investigated whether phonetic retuning would take place between different talkers who spoke with a foreign (i.e., Dutch) accent. They found that the perceptual retuning induced by the female talker generalized to another female talker but to a male talker only if his speech sounds were sampled at a similar perceptual span to the female's. The results suggested that cross-talker generalization occurs when two talker's acoustics exhibit a certain level of similarity.

Likewise, using a variant of the lexically guided perceptual learning paradigm Xie, Theodore and Myers (2017) exposed native English-speaking subjects to Mandarin-accented speech produced by a single talker or multiple talkers (Xie & Myers, 2017) and then tested if learning generalized to another Mandarin accented talker. In these studies, the critical trials during the exposure stage contained Mandarin-accented /d/-ending words (e.g., "alongside"). Mandarin-accented /d/- and /t/-ending English words are ambiguous for native English speakers since the perceptual cue that most reliably distinguishes the contrast is burst duration instead of vowel duration which English speakers use for natively accented /d/ and /t/. Exposure conditions in which listeners were exposed to the ambiguous Mandarin productions were compared to control conditions in which listeners did not hear any /d/-ending words. To assess adaptation at test, an online measure of lexical processing was used (cross-modal priming). The results showed that exposure to the ambiguous acoustic-phonetic characteristics of Mandarin-accented speech facilitated subsequent lexical processing. Learning generalized to new utterances from the same speaker, which aligned with the findings from studies that employed the original paradigm of lexically guided perceptual learning. However, although generalization of learning was observed between the two different talkers who were acoustically similar, it was not observed between two talkers who were acoustically dissimilar (shown in Figure 1). More importantly, perceptual adaptation and generalization showed a graded effect in terms of inter-talker similarity such that facilitation of lexical processing was largest for the same speaker, robust but reduced for the similar talker, and absent for the dissimilar talker. These researchers thus argued that cross-talker generalization depends on the acoustic similarity between the two talkers regardless of whether there is shared group-based variation. If the two talkers were sufficiently similar, successful generalization would occur.

Xie, Liu and Jaeger (2021) replicated the study described above by Bradlow and Bent (2008), and they argued that generalization is talker-to-talker and similarity-based rather than group-based. After comparing the acoustic similarity between every exposure talker to the test talker, they claimed that generalization after exposure to multiple talkers depends on the similarity between particular exposure talkers and the test talker. More specifically, whether the generalization could turn out to be successful depended on whether there was a talker, among all the exposure talkers, that exhibited high levels of similarity with the test talker.

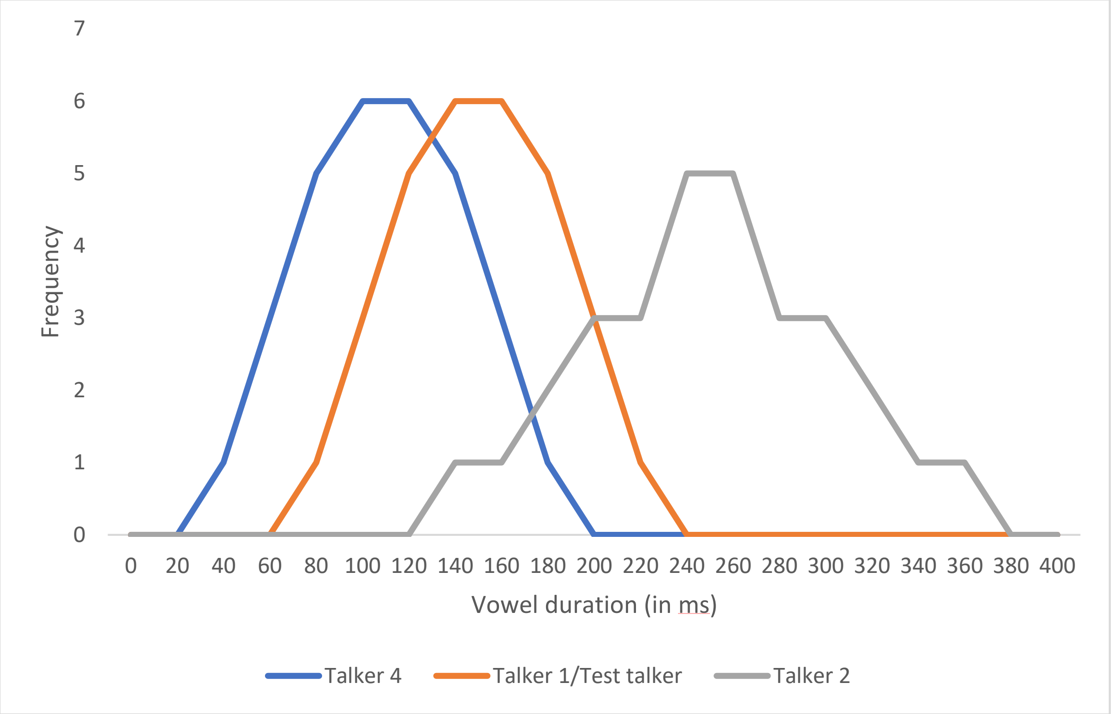


Figure 1. Schematic illustration of the talker's acoustic similarity in Xie & Myers (2017). The y-axis depicts frequency, and the x-axis depicts vowel durations of particular /d/-final productions. Talker 4 is the "similar" talker, and Talker 2 is the "dissimilar" one.

Predictions from theoretical models for speech adaptation

The findings from Xie and Myers (2017) and others showing the importance of acoustic similarity in perceptual adaptation are generally consistent with exemplar-based models of speech perception (Goldinger, 1998; Johnson, 2006). According to these models, fine-grained acoustic details, particularly those associated with talker-related factors, are encoded in episodic memory for spoken language. A memory trace, therefore, of each spoken word or individual speech sound that is encountered, is formed including indexical and other properties of spoken language such as speaking rate, vocal effort, and intonation contour (Bradlow, Nygaard & Pisoni, 1999; Church, & Schacter 1994; Goldinger, 1996; Kapnoula & Samuel, 2019). When words are recognized, the incoming perceptual signal is compared to multiple exemplars in the mental lexicon and categorized based on similarity to the accumulated memory traces. Thus, the model easily accounts for talker specificity in speech perception. Learning and generalization are based on similarity in indexical and linguistic form to the listeners' cumulative memory representations of spoken language. Even though Xie and colleagues did not specify the role that memory plays in cross-talker generalization, exemplar models claim that listeners could generate specific representations for every talker encounter and compare the similarity for generalization.

Talker-specificity effects and reliance on talker similarity are consistent not only with exemplar-based accounts but also with the Bayesian belief-updating models. The ideal adaptor model, proposed by Kleinschmidt and Jaeger (2015), provided a computational account for both talker-specific and talker-independent representations in speech perception. The ideal adaptor model generally treats speech perception as a problem of cue-to-category mapping with probabilistic relationships, which requires tracking statistical properties (i.e., mean and variance; Clayards, Tanenhaus, Aslin & Jacobs, 2008) and updating knowledge (i.e., the priors; Feldman, Griffiths & Morgan, 2009; Norris & McQueen, 2008). Essentially, the model engages in distributional learning (Feldman, Griffiths, Goldwater & Morgan, 2013; McMurray, Aslin & Toscano, 2009; Vallabha, McClelland, Pons, Werker & Amano, 2007) when encountering a novel situation (e.g., a single talker or a group of accented talkers) for which it creates new, or adjusts already existing, categories. Therefore, between the acoustic level and the lexical level, a well-trained ideal adaptor contains a hierarchical structure where distributions of talker-specific acoustic features are represented for each individual talker (also called "talker model") while group-level distributions ("group models") are represented at a higher level.

Although previous findings are largely consistent with extant models of speech perception such as exemplar-based and Bayesian ideal adaptor accounts in which listeners are able to distributionally track acoustic variation produced by individual talkers and generate talker-specific representations of speech, it remains unclear how cumulative, and potentially long-term, previous experiences influence perceptual adaptation and cross-talker generalization (i.e., prior category structure and how a change in that structure occurs during exposure). Therefore, it is of interest to ask whether and what type of prior experience influences how perceptual mechanisms are tuned during the perceptual learning of spoken language.

Directionality in speech adaptation

The relationship between the stimuli presented in the learning and test conditions in lexically guided perceptual learning paradigms leads to the question of whether providing exposure to atypical pronunciations that are progressively more or less similar to native category structures would impact talker-specific adaptation and generalization. Sumner (2011) investigated perceptual adaptation to French-accented English by examining whether the organization of stimulus materials during exposure and learning would influence the degree of adaptation.

Critically, during learning, stimuli were presented from native to accented productions or from accented to native productions. The results showed that categorization boundaries signaling perceptual learning shifted more towards the acoustic-phonetic space marked by the accented speech if the stimuli were presented in an order that started with the native stimuli and transitioned to progressively more accented productions.

Maye, Aslin, and Tanenhaus (2008) studied this particular question by exposing subjects to words in which the critical segment was shifted downward or upward in formant space (i.e., shifting high vowel /i/ down to mid vowel /ɛ/ or vice versa). The results showed that subjects identified nonwords that had the vowel downward-shifted as real words but did not identify nonwords that had the vowel upward-shifted as real words. The study suggests that perceptual learning depends on and is specific to the particular acoustic-phonetic space. Perceptual adaptation can be directionally-specific and constrained by linguistic knowledge. Likewise, Melguy and Johnson (2022) provided the evidence that supports the claim that what is behind boundary shifts during perceptual adaptation is a non-uniform and specific category expansion. Using the lexically guided perceptual learning paradigm, when subjects were exposed to /θ/-biasing stimuli midway between /θ/ and /s/ category, a category boundary shift reflecting perceptual learning was observed for a /θ/-/s/ continuum. More critically in this study, generalization of the boundary shift to a continuum formed with a novel phonetic category is constrained by to which the direction the boundary is shifted. Given the same /θ/-biasing exposure, such boundary shift was generalized to the /θ/-/ʃ/ continuum in which the novel phonetic category /ʃ/ is in the direction to which the boundary is shifted but not to the /θ/-/f/ continuum in which the novel category /f/ is on the opposite direction.

The current study

Given that both talker similarity and prior linguistic category knowledge appear to modulate generalization of perceptual adaptation and learning of atypical phonetic categories (e.g., accented speech), we hypothesize that prior experience and directionality in acoustic-phonetic space would impact cross-talker generalization. If cross-talker generalization depends solely on similarity comparison as suggested by Xie and Myers (2017), then in which direction listeners generalize from one talker to the other talker should NOT impact whether listeners will generalize. For example, among the exposure talkers presented in their study (schematically shown in Figure 1), although Talker 2 is more acoustically different from the test talker, this talker's vowel duration is quite similar to native American English speakers. That is, being exposed to talker 2's acoustics would likely resemble having exposure to a native speaker whose vowel duration clearly falls into the /d/ category. The accuracy of critical words during the exposure phase in Xie and Myers (2017) also favors the claim. Listeners showed higher accuracy in judging whether an item was a word when hearing stimuli recorded from Talker 2 (0.84, more English-like) than from Talker 4 (0.72, less English-like), suggesting that Talker 2's pronunciations were less ambiguous for the listeners.

Experiment 2 by Xie and Myers (2017) exhibited a situation in that subjects were exposed to a talker with a slight Mandarin accent (Talker 2) and show no generalization to a test talker (Talker 1), and they attributed such result to a lack of acoustic similarity between the two talkers. However, such a result can also be explained by lexically guided structural change in phonetic categories. Figure 2A hypothetically illustrates how the category would be changed after being exposed to an accent that is less Mandarin but more English-like. The change that occurs in the listener's category structure results in almost no categorization boundary shift and is formally predicted by an ideal adaptor (Kleinschmidt & Jaeger, 2015). In order to tease apart the two explanations, the exposure talker and test talker need to be simply switched. As shown in Figure 2B, the listener hears a typical Mandarin-accented speaker and then generalize to a nearly native English-accented speaker, even though the direction of generalization becomes the opposite (namely, from less to more accented to from more to less accented), the acoustic similarity is the same.

The current study is designed to investigate whether cross-talker generalization is based exclusively on similarity in acoustic-phonetic features between individual talkers or alternatively influenced by underlying changes in phonetic categories in which experimental exposure and past perceptual experience both play a role. I use the same experimental procedure adopted by

Xie and Myer (2017) with a lexical decision task at exposure and a cross-modal priming task for the test. An anonymous Mandarin speaker was recorded to serve as the test talker. The exposure "talker" was created from the original utterances so that ONLY the vowel durations associated with the /d/ productions for all the critical trials in both exposure and test, were extended such that they are 80 ms longer than the original. This manipulation allowed for precise control of similarity across "talkers" and created stimulus sets that were more (talker + 20ms) or less (original talker) similar to native English phonetic categories.

The two sets of stimuli then served as the test talker (unmodified-Talker 1) and the exposure talker (with 80 ms extension-Talker 2). In Experiment 1, listeners were exposed to and tested on the same talker (i.e., Talker 1 to Talker 1). The same talker conditions were designed to verify the validity of the tasks for the following experiments. Critically, Experiment 2 creates an exposure-generalization scenario in which subjects will be exposed to different talkers to assess the conditions under which cross-talker generalization occurs. In one situation, listeners will be exposed to Talker 2 and then tested on Talker 1 (Experiment 2a), namely, in a direction from less Mandarin-accented to more Mandarin-accented (see Figure 2A). We do not expect to see generalization in this condition. In the second situation (Experiment 2b), listeners first heard ambiguous stimuli (typically Mandarin-accented) from Talker 1 and then tested with Talker 2 who produces unambiguous stimuli (less Mandarin-accented but more like native English-accented; see Figure 2B), In Experiment 2b, because the /d/-category would be expected to expand towards the ambiguous space, generalization of learning would be expected. If such a pattern is observed, it would suggest that the underlying and pre-existing phonetic category structures, namely cumulative previous perceptual experiences, contribute to the degree to which listeners perceptually adapt to atypical or accented pronunciation because acoustic similarity between the two talkers was held constant across two experiments in Experiment 2.

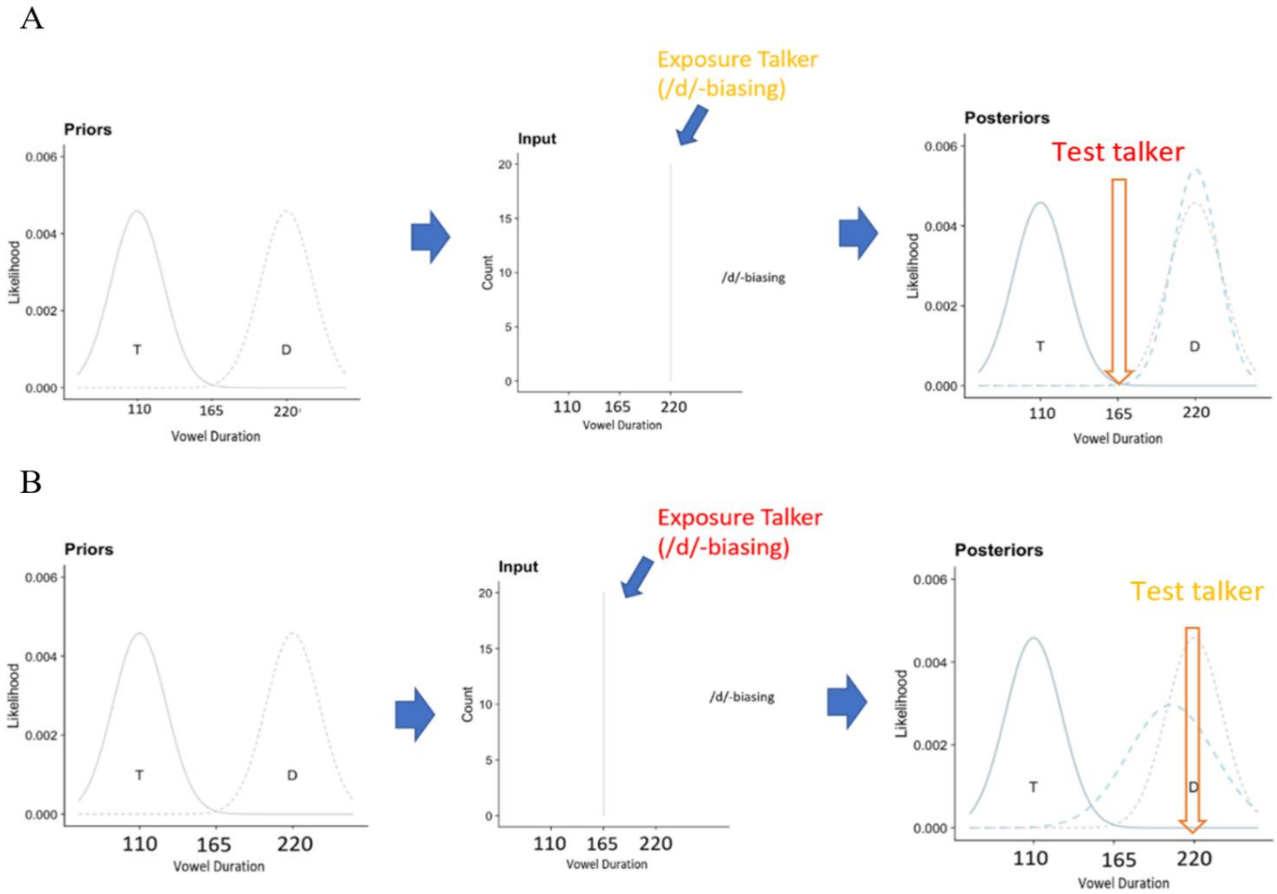


Figure 2. Experimental procedures inspired by Xie and Myers (2017) with the ideal adaptor's predictions of how category would change depending on the ambiguity (Mandarin-accentedness) of the exposure stimuli. A corresponds to Experiment 2a, and B corresponds to Experiment 2b.

Experiment 1

Experiment 1 was designed to replicate and extend Xie, Theodore and Myers (2017). Listeners were exposed to a Mandarin-accented talker producing ambiguous /d/-final words and were tested on their degree of perceptual adaptation in a cross-modal priming task to the same talker. The ambiguous accented /d/-final words were presented in lexically disambiguating contexts in the experimental condition and replaced with words without /d/-final segments in the control condition. The cross-modal priming task was used at test to assess the effects of learning on online lexical processing. In this test task, the critical trials consisted of auditory /d/-final word primes presented with visual /d/ and /t/ final target words (e.g., seed-SEED and seed-SEAT) and were compared to unrelated prime-target pairs (e.g., fair-SEED and fair-SEAT). Priming effects were calculated as the benefit in response time for related relative to unrelated pairs; RTs to unrelated words minus RTs for related words (e.g., fair-SEED minus seed-SEED). The related primes are expected to induce faster RTs than the unrelated primes, therefore, we expected positive values for the priming effect. If listeners perceptually tune to the atypical /d/-final productions, then the experimental group who received exposure to those productions should have a larger priming effect for the /d/-final prime pairs than the control group who did not receive exposure to the critical /d/-final words.

Participants

93 monolingual English speakers (exposure: 51, control: 42) with no hearing and visual disorders were recruited through Prolific (https://www.prolific.co), an online platform that provides a participant-recruiting interface. Participants were pre-screened to ensure that they were raised in an English-speaking monolingual environment. Participants were randomly assigned to the exposure groups (experimental vs. control). Before the experiment, participants were given online informed consent according to Emory University Institutional Review Board. Participants were paid $8/hour for their participation. 27 subjects (exposure: 14, control: 13) were excluded for taking more than five minutes between exposure phase and test phase, not passing the headphone check, self-reporting as bilingual speakers or having exposure to Mandarin or having had or having speech or hearing disorders, or not finishing the study. 66

subjects were included in the following analyses with 37 in the experimental condition and 29 in the control condition.

Materials

Speech materials were recorded by a 23-year-old male native Mandarin speaker. The speaker started taking English courses at 8 years old but did not report native-level exposure and use of English. He had resided in the United States for 15 months at the time of recording. Digital recordings were made in a sound-attenuated room using a Samson C01U Pro microphone onto a MacBook Pro running Audacity sound-editing software. Stimuli were digitally sampled at 44.1 kHz and amplitude normalized. Figure 3 shows probability density functions of vowel durations of critical exposure (left panel) and test words (right panel). Vowel duration is important here because it is a cue on which English speakers rely to make a distinction between /d/ and /t/. The natural vowel durations were measured such that the onset is identified at the zero-crossing when the waveform of the vowel becomes consistent and starts reoccurring at the same size, and the offset is identified at the zero-crossing when the waveform of the vowel becomes weak and stops reoccurring at the same size. Onset and offset locations were confirmed on accompanying spectrograms showing the beginning and end of clear periodicity. The acoustic measurements were done using Praat (Boersma, & Weenink, 2023). These measurements confirmed that the speaker recorded for this study produced English words with a typical Mandarin accent, comparable to the acoustic patterns reported for the stimuli used in Xie, Theodore and Myers (2017) and Xie and Myers (2017). Of note, the speaker’s /t/-final productions were recorded for comparison (right panel Figure 3). Subjects only heard /d/-final words during test.

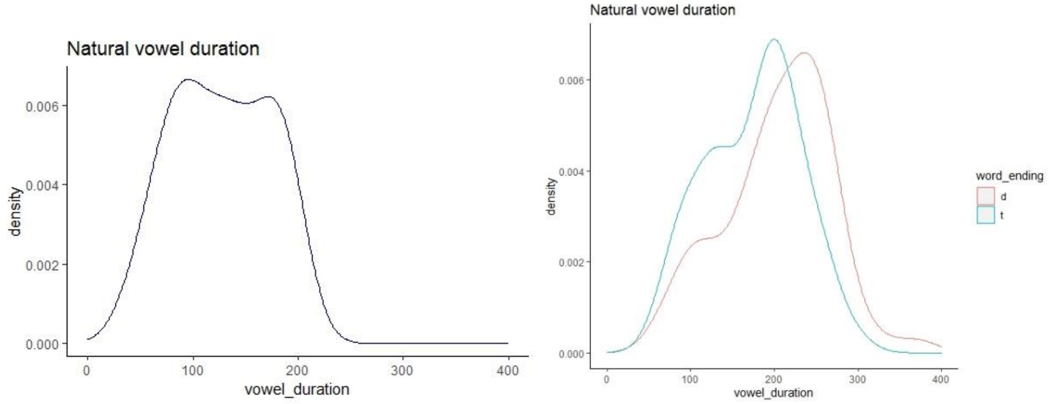


Figure 3. The graphs show probability density functions of critical words in exposure (left panel) and test words (right panel) in terms of their vowel durations in milliseconds.

Exposure

There were two conditions for the exposure phase, experimental and control. The wordlist consisted of 180 items in total. The 180 items included 60 filler words, 90 nonwords and 30 critical words for the experimental group or 30 replacement words for the control group. The critical words were words that ended with the stop consonant /d/ (e.g., alongside). The replacement words (e.g., animal) were matched to the critical words in syllabic length and mean lemma frequency in CELEX (Baayen, Piepenbrock & Guliker, 1995). All items in the wordlist contained three to four syllables. Since perceptual learning in the study should mainly be induced by alveolar stops (e.g., /d/), in order to avoid possible influence introduced by other consonants, words were selected to meet the following criteria: (a) /d/ appeared only in wor/d/-final position, and only in critical words; (b) no other alveolar stops, no other voiced stops or dental fricatives, and no postalveolar affricates occurred; and (c) no voiceless stops (/p/ or /k/) occurred in wor/d/-final position. The test stimuli were selected using the same criteria.

Test

Both exposure groups heard the same list of items at test, which consisted of 60 monosyllabic /d/-final words (each of which has a /t/-final minimal pair, e.g., *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, and did not differ, $t(59) = .159$, $p = .88$.

Procedure

Exposure

Subjects were asked to complete a lexical decision task during exposure followed by a cross-modal priming task during test. During exposure, the experimental group heard the experimental list (e.g., critical /d/-final words) whereas the control group heard the control list (e.g., replacement words). The items were played in a random order. For every trial, subjects heard an item and then were asked to judge whether it was a word or not by pressing corresponding keys (i.e., A and L) on their keyboards. The assignment of “yes” and “no” responses to the keys was counterbalanced across subjects.

Test

The test phase was the same for both groups. Subjects were told that they would hear an item and then see another item printed on the screen at the offset of the auditory stimuli. They were asked to respond with a key press (“yes” and “no” response mappings to the keys A and L counterbalanced as in the exposure phase) to judge whether the item they saw was a word or not. 60 critical trials were comprised of four different prime-target (auditory-visual) pairing types; identity prime (e.g., seed-SEED), unrelated prime (/d/-final word as visual target, e.g., fair-SEED), minimal pair (e.g., seed-SEAT) and unrelated prime (/t/-final word as visual target, e.g., fair-SEAT). Words in each set of minimal pair items were rotated over four counterbalanced lists. Within each list, there were equal proportions of the four pairing types (15 trials for each). The rest of the trials were non-critical and identical across counterbalanced lists. 15 filler words were paired with an identical prime (e.g., same–SAME) and another 15 with an unrelated prime (e.g., care–ROAR). Therefore, there were 90 trials that should have been correctly responded to with "yes", the printed item was a word. Another 90 auditory filler words were paired with visual nonwords (e.g., sleeve–OURN) that should have been correctly responded to with "no". The wordlist was pseudorandomly arranged such that four consecutive trials requiring the same response would not occur. The randomized order was also counterbalanced by presenting two reverse orders.

Since the study was conducted online using Gorilla Experiment Builder (www.gorilla.sc), subjects were instructed to complete the experiment using their desktops (phones and tablets disallowed). They were also reminded to wear headphones and adjust the volume to 50% when hearing the stimuli during practice. There was a headphone check before the actual experiment started in which subjects were instructed to judge which one of the three tones played is the softest (Siegel, Traer & McDermott, 2017). All the information regarding their devices was registered automatically by Gorilla Experiment Builder. Before both exposure and test phases, ten practice trials were given to the subjects to familiarize them with the tasks.

Figure 4 shows the order of presentation of stimuli on each trial for the lexical decision task (three upper panels) used during the exposure phase and the cross-modal priming task (three lower panels) used during the test phase. For the lexical decision task, the fixation appeared and stayed on for 1000ms (the first screen, top left panel). At the offset of the fixation, the stimulus was played as the prompt appeared (the second screen, top middle panel). At the offset of the sound, subjects were expected to respond according to the prompt by key press (the third screen, top right), the screen would last for 3000ms or until a response was made. For the cross-modal priming task, after the fixation appeared and stayed on for 1400ms (the first screen, lower left panel), the stimulus was played as the prompt appeared (the second screen, lower middle panel). At the offset of the sound and fixation, subjects were expected to respond to the printed word according to the prompt by key press (the third screen, lower right), the screen would last for 2000ms or until a response was made. All the items that were presented across two screens (i.e., the prompt in the lexical decision task and the fixation in the cross-modal priming task) were placed at the exact same location on the screen.

At the end of the experiment, subjects were asked to fill out the form regarding their language background. Also, they were asked whether they thought the speaker at exposure and at test were the same person or not.

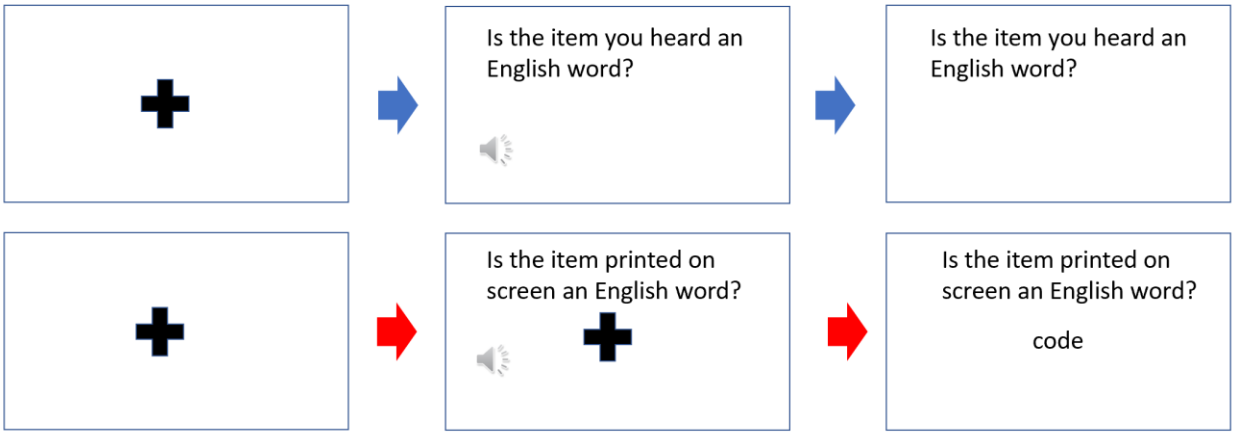

Is the item you heard an English word?
Is the item you heard an English word?
Is the item printed on screen an English word?
Is the item printed on screen an English word?
code

Figure 4. The figure shows the presentation of stimuli during a single trial for the lexical decision task during exposure (three upper panels) and the cross-modal priming task during test (three lower panels).

Results

Exposure

Response accuracy reflects how many /d/-final critical words were correctly judged as real words in the experimental group (M_{exp} = 0.83, SD_{exp}= 0.13) and how many replacement words were correctly judged as real words in the control group (M_{con} = 0.73, SD_{con} = 0.09). Response accuracy was numerically higher for the critical words than for the replacement words. Although there is a difference in response accuracy, both the experimental and control performance was well above chance, and subjects performed well for the critical /d/-final items. Response accuracy for filler words (M_{exp} = 0.82, SD_{exp} = 0.09; M_{con} = 0.84, SD_{con} = 0.06) and for nonwords (M_{exp} = 0.91, SD_{exp} = 0.06; M_{con} = 0.88, SD_{con} = 0.15) was comparable across conditions.

Test

Table 1 shows mean response time (RT) across participants in the test phase. RTs of correct responses only were analyzed. Four words (*spate, moot, plod, pleat*) were discarded due to lower than chance level accuracy and trials that were not responded to because of technical errors were excluded. Additionally, responses (5.07% of correct trials) above or below 2 SDs from the mean of each prime type in each group were excluded from the RT analysis.

A repeated measures ANOVA was conducted with RT as the dependent measure and exposure group (experiment vs. control), prime type (related vs. unrelated) and target type (/d/-final vs. /t/-final) as factors. A significant main effect of prime type was found ($F(1, 64) = 38.645$, $p < 0.001$). RTs for the related prime were faster than for the unrelated prime. In addition, a significant main effect of target type was found with RTs for the /d/-final target slower than for the /t/-final target ($F(1, 64) = 16.541$, $p < 0.001$). No other main effects or interactions were significant at the p < 0.05 level, including the three-way interaction between exposure group, prime type and target type (*F*(1, 64) = 0.571, *p* = 0.453). Figure 5 illustrates the results of RTs in the three-way interaction.

Table 1

Mean error rates and Reaction Time (RT) across participants in the cross-modal priming task as a function of exposure group in Experiment 1

Mean % error	/d/-final		/t/-final	
	related prime	unrelated prime	related prime	unrelated prime
Experimental	5 (7)	1 (11)	3 (6)	6 (7)
Control	5 (6)	13 (18)	4 (7)	8 (14)
RT (milliseconds)	/d/-final		/t/-final	
	related prime	unrelated prime	related prime	unrelated prime
Experimental	650 (130)	672 (102)	626 (150)	647 (100)
Control	638 (129)	673 (135)	600 (142)	639 (105)

Note. SD is provided in parentheses.

Experiment 1

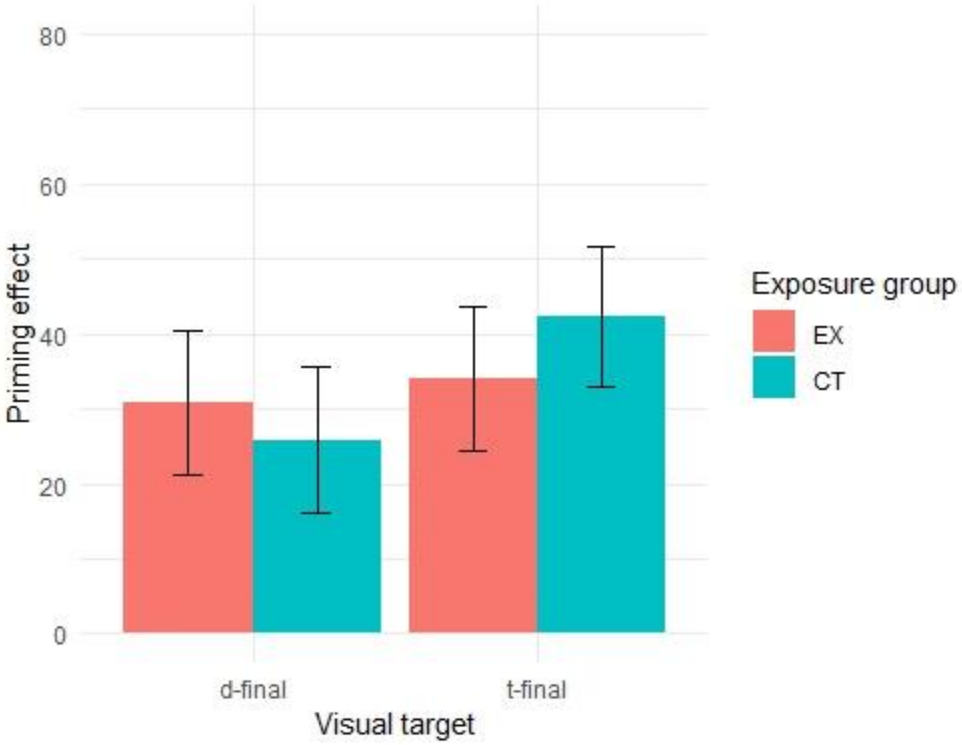


Figure 5. Experiment 1: Priming of /d/-final words (reaction time [RT] in fair-SEED trials minus RT in seed-SEED trials) and /t/-final words (RT in fair-SEAT trials minus RT in seed-SEAT trials) for participants exposed to critical words (experimental group) or replacement words (control group). Error bars represent standard error of the mean.

Discussion

The pattern of results did not replicate the findings in Xie, Theodore and Myers (2017), which was contrary to our expectations. Although there were a number of differences in the implementation of the two studies, one possible explanation for this failure to replicate is that the talker who was recorded for the current study may have differed in specific ways from the talker in the Xie et al study. Both talkers were male native Mandarin speakers, but the talker used in this study produced the multisyllabic (critical words in exposure phase) and monosyllabic words (critical words at test) with shorter and longer vowel durations respectively, while the talker in Xie et al's study produced similar closure durations across exposure and test. Figure 3 shows the acoustic distribution of the two sets of words used during exposure and test for the talker used in the current study. We performed a Welch two sample t-test comparing the vowel durations of critical exposure words and of critical test words (/d/-final). The result shows that the two distributions are significantly different ($t(76.11) = 6.64$, $p < 0.001$), indicating that the two distributions are different. Although Xie, Theodore and Myers (2017) did not report whether their critical exposure words were acoustically aligned with their critical test words, the two acoustic distributions representing vowel duration in their study look quite similar to one another, suggesting that their speaker produced monosyllabic and multisyllabic words in a similar range in contrast to our current stimulus materials. There are a number of reasons why there might be relative differences or not in exposure and test words. Because vowel duration depends on speaking rate, speakers in either study could have produced exposure and test words at different rates, changing the correspondence in vowel duration across contexts and studies.

Given the results, the original finding that exposure to Mandarin-accented productions would facilitate generalization to new words was not supported. Rather, the results indicated that the acoustic differential in vowel duration between the critical words in exposure and test may have hindered generalization. Interestingly, because in the current study, the speaker's vowel durations in the critical exposure words were more typical of Mandarin-accented productions whereas the vowel durations in the test words were more typical of native English productions, the misalignment happened to create a situation which we originally planned to test in Experiment 2, with exposure to Mandarin-accented vowel durations and being tested on more native-like English durations. This misalignment and the direction of the misalignment may have blocked generalization of learning.

In Experiment 2a, we implemented an 80 ms extension of vowel duration for critical words with (Talker 2), shifting the distribution to be more acoustically similar to the naturally recorded critical test words (Talker 1). The two distributions were in fact similar ($t(76.11) = -0.224$, $p = 0.82$) and therefore should result in successful generalization. For Experiment 2b, the 80 ms shift in distribution implemented for the test words made the acoustic differences across exposure and test the largest among the three experiments and thus, we should not observe generalization of learning, since the distributions of critical exposure and test words were further apart in the acoustic space. Of note, here we are only re-stating the original manipulation, we are NOT adding any new manipulations. However, we are now considering a variation which we previously assumed having no effect to the results and would like to see whether the data that would be collect in Experiment 2 would support our new predictions.

Experiment 2

Experiment 2 was designed to examine the extent to which listeners are sensitive to the relative similarity in acoustic properties of a talker's accented productions. Two questions were addressed. The first was whether listeners would generalize perceptual learning to a "talker" whose distribution of acoustic cues (vowel duration) for the critical segment /d/ was altered. Using the same paradigm as in Experiment 1, listeners were exposed to accented /d/-final utterances from one distribution, the same ones used in Experiment 1, and then tested on d/-final utterances from a artificially manipulated distribution (mean vowel duration shifted 80 ms) in Experiment 2a, and listeners were exposed to artificially manipulated accented /d/-final utterances (mean vowel duration shifted 80 ms) and then tested on d/-final utterances, the same ones used in Experiment 1 (see Figure 6) All other aspects of the exposure and test "talker" remained identical. Two "talkers" were different for the critical acoustic dimension, but all other characteristics of the productions were held constant. The second question was whether generalization of perceptual learning depended on similarity to native English phonetic categories. Again, the two "talkers" were different for the critical acoustic dimensions, but across conditions the exposure talker either produced stimuli that were less Mandarin-accented but more natively English-accented relative to the test talker or the exposure talker produced stimuli that were more Mandarin-accented but less natively English-accented relative to the test talker.

For experiment 2a, listeners heard critical exposure words in which the distribution of vowel durations was shifted 80 ms to longer values (an 80 ms extension, Talker 2) and were tested on the words that are identical to the ones used in Experiment 1 (Talker 1). For experiment 2b, listeners heard exposure words that were identical to the ones used in Experiment 1 (Talker 1) and were tested on critical trials in which the distribution of vowel durations was shifted by 80 ms (Talker 2).

Although the original rationale for Experiment 2 is outlined above, since Experiment 1 did not yield any significant results, we evaluated the extent to which an acoustic misalignment in vowel duration between the critical exposure words and critical trials at test may have been responsible for the failure to replicate the Xie and Myers (2017) learning effect. Thus, in Experiment 2a, if misalignment in vowel duration accounted for the null results in the first experiment, we hypothesized that with the 80 ms extension in the critical exposure words, the words at exposure would be more acoustically aligned with the auditory words at test such that learning as instantiated by a larger priming effect should be observed for the experimental group and not for the control group. In Experiment 2b, however, because the 80 ms extension was made on the critical test trials, the critical exposure words and the auditory words presented at test would be more misaligned acoustically. Therefore, we expected no effect of learning between the two exposure groups.

Participant

85 monolingual English speakers (exposure: 44, control: 41) were recruited through Prolific (https://www.prolific.co) for Experiment 2a and 91 (exposure: 46, control: 45) for Experiment 2b. Participants were pre-screened to ensure that they were raised in an English-speaking monolingual environment. Participants were randomly assigned to the exposure groups (experimental vs. control). Before the experiments, participants were given online informed consent according to Emory University Institutional Review Board. Participants were paid $8/hour for their completion. In Experiment 2a, 24 subjects (exposure: 11, control: 12) were excluded for taking more than five minutes between exposure phase and test phase, not passing the headphone check, self-reporting as bilingual speakers or having exposure to Mandarin or having had or having speech or hearing disorders, or not finishing the study. In Experiment 2b, 41 subjects (exposure: 22, control: 19) were excluded for the same reasons. For experiment 2a, 62 subjects were included in the following analyses with 33 in the experimental condition and 29

in the control condition. For experiment 2b, 50 subjects were included in the following analyses with 24 in the experimental condition and 26 in the control condition.

Materials

The difference between Experiment 2a and 1 is that the stimuli used during exposure phase were manipulated such that every critical word had an 80-milliesecond extension in their vowel duration. The difference between Experiment 2b and 1 is that the stimuli used during test phase were manipulated such that every /d/-final word had an 80-milliesecond extension in their vowel duration. The acoustic manipulations were made using the original stimuli from Experiment 1. The acoustic extension was made based on the measurements described in Experiment 1 through Praat (Boersma & Weenink, 2023). Figure 6 shows the distribution of acoustic values of critical words used in Experiment 2. In Figure 6a, the red line represents the acoustic distribution of critical words used in experiment 1 and the green line represents the acoustic distribution of critical words used in experiment 2a. In Figure 6b, the red line represents the acoustic distribution of /d/-final words used in experiment 1 and the green line represents the acoustic distribution of /d/-final words used in experiment 2b.

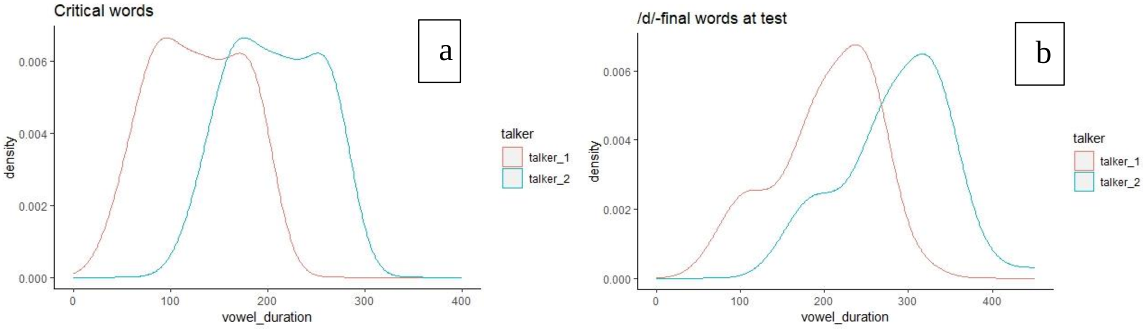


Figure 6. This figure shows a) the distribution of acoustic values of critical words (at exposure) used in Experiment 1 (Talker 1) and 2a (Talker 2); the stimuli used at test between the two experiments are identical and b) shows the distribution of acoustic values of /d/-final words (at

test) used in Experiment 1 (Talker 1) and Experiment 2b (Talker 3); the stimuli used at exposure are identical.

Procedure

The procedure was the same as in Experiment 1.

Results

Exposure

The pattern of response accuracy in Experiment 2 was similar to that found for Experiment 1. In Experiment 2a, response accuracy for critical words indicated how many /d/-final words were correctly judged as real words by the subjects in the experimental group (M_{exp} = 0.85, SD_{exp} = 0.12) and how many replacement words were correctly judged as real words in the control group (M_{con} = 0.70, SD_{con} = 0.10). Response accuracy for filler words (M_{exp} = 0.85, SD_{exp}= 0.08; M_{con}= 0.83, SD_{con}= 0.09) and for nonwords (M_{exp} = 0.88, SD_{exp} = 0.14; M_{con}= 0.89, SD_{con} = 0.11) across conditions was comparable. In Experiment 2b, response accuracy for critical words in the experimental group (M_{exp} = 0.83, SD_{exp} = 0.11) was also higher than for replacement words (M_{con} = 0.69, SD_{con} = 0.11). Response accuracy for filler words (M_{exp} = 0.82, SD_{exp} = 0.10; M_{con} = 0.85, SD_{con} = 0.07) and for nonwords (M_{exp} = 0.90, SD_{exp} = 0.05; M_{con} = 0.92, SD_{con} = 0.06) was comparable across conditions. Across Experiment 2a and 2b, response accuracy was numerically higher for the critical words than for the replacement words. Although there is a difference in response accuracy, both the experimental and control performance was well above chance, and subjects performed well for the critical /d/-final items.

Test

Table 2 shows mean RT across participants in the test phase for Experiment 2a and 2b. RTs of correct responses only were analyzed. Four words (*spate, moot, plod, pleat*) were discarded due to lower than chance level accuracy and trials were not responded to because of technical errors were excluded. Additionally, responses (5.33% of correct trials in Experiment 2a and 4.44% in Experiment 2b) above or below 2 SDs from the mean of each prime type in each group were excluded from the RT analysis.

As in Experiment 1, repeated measures ANOVAs were conducted with RT as the dependent measure and exposure group (experiment vs. control), prime type (related vs. unrelated) and target type (/d/-final vs. /t/-final) as factors for both Experiment 2a and 2b. In Experiment 2a, a significant main effect of prime type was found (F (1, 60) = 55.013, $p < 0.001$) with faster RTs for the related prime than for the unrelated prime. In addition, a significant main effect of target type was found with RTs for the /d/-final target slower than for the /t/-final target (F(1, 60) = 8.569, $p < 0.01$). No other main effects or interactions were significant at the p < 0.05 level, including the three-way interaction between exposure group, prime type and target type (F(1, 60) = 1.157, $p = 0.283$).

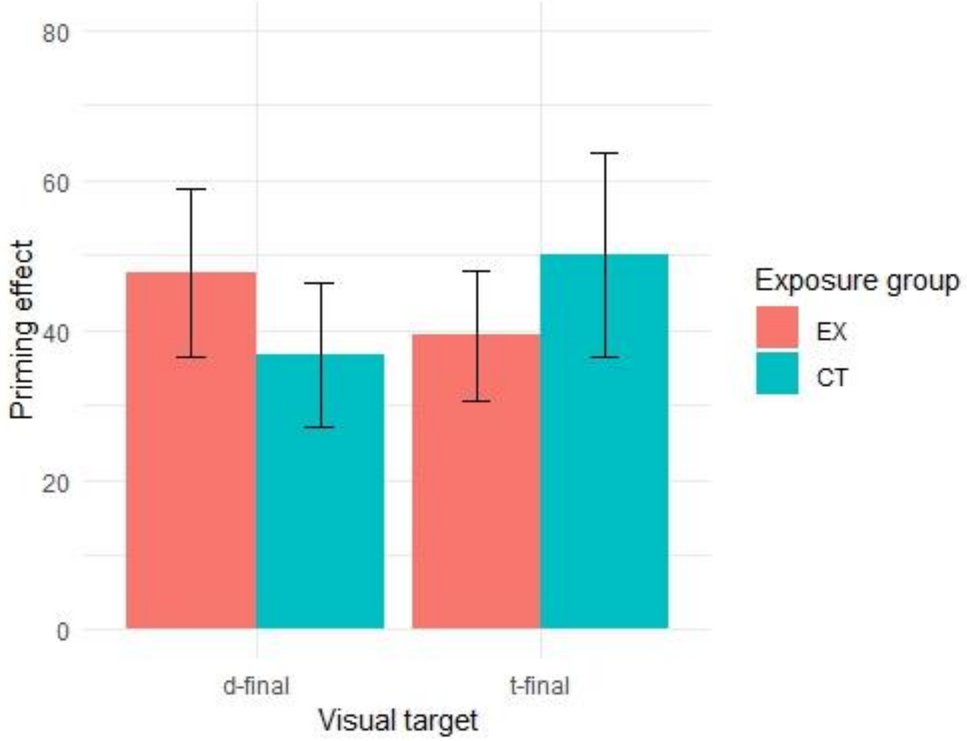
For Experiment 2b, a significant main effect of prime type was found (F (1, 48) = 78.836, $p < 0.001$) with faster RTs for the related prime than for the unrelated prime. A significant main effect of target type was also found with RTs for the /d/-final target slower than for the /t/-final target (F(1, 48) = 8.964, $p < 0.01$). In addition, there was a significant interaction between prime type and target type (F(1, 48) = 5.96, $p = 0.018$), suggesting a larger priming effect overall for the /d/-final words as compared to the /t/-final words. No other main effects or interactions were significant at the p < 0.05 level, including the three-way interaction between exposure group, prime type and target type (F(1, 48) = 0.416, $p = 0.522$). Figure 7 illustrates priming as a function of target type and condition in Experiment 2.

Table 2

Mean error rates and Reaction Time (RT) Across Participants in the Cross-Modal Priming Task as a Function of Exposure group in Experiment 2

Mean % error	/d/-final		/t/-final	
2a	related prime	Unrelated prime	related prime	unrelated prime
Experimental	4 (5)	7 (8)	3 (6)	2 (4)
Control	6 (8)	8 (9)	4 (6)	4 (6)
2b				
Experimental	7 (9)	11 (10)	4 (6)	9 (9)
Control	6 (8)	10 (10)	4 (7)	6 (10)
RT (milliseconds)	/d/-final		/t/-final	
2a	related prime	unrelated prime	related prime	unrelated prime
Experimental	671 (150)	720 (163)	633 (130)	675 (134)
Control	642 (143)	677 (116)	613 (128)	658 (127)
2b				
Experimental	669 (152)	733 (152	648 (117)	686 (101)
Control	611 (77)	680 (79)	618 (78)	653 (94)

Experiment 2a



Experiment 2b

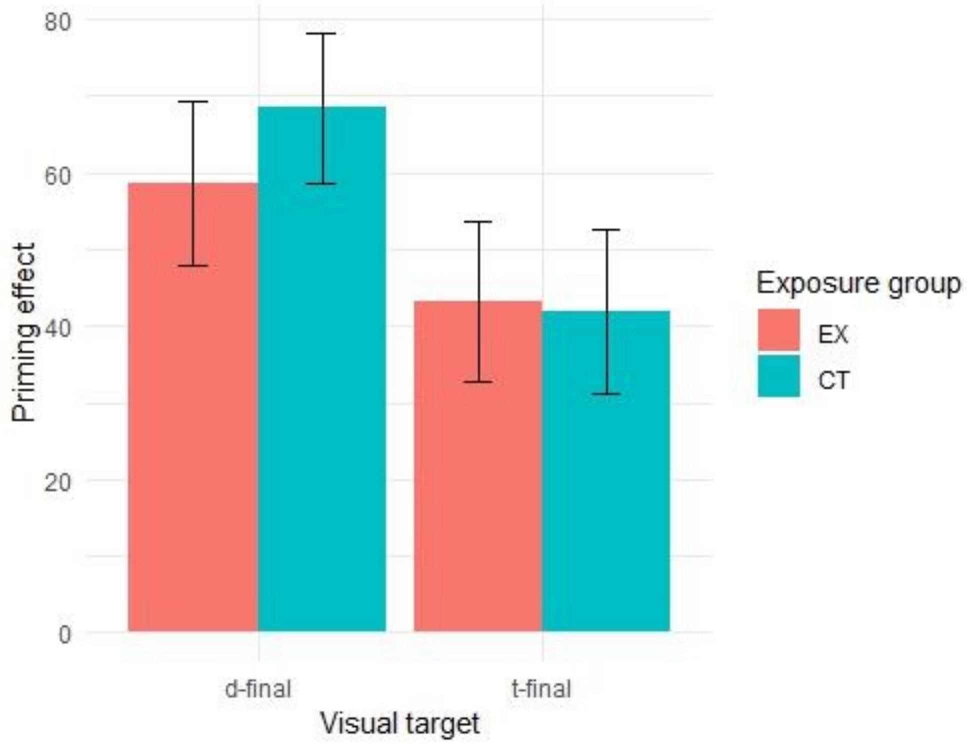


Figure 7. Experiment 2: Priming of /d/-final words (reaction time [RT] in fair-SEED trials minus RT in seed-SEED trials) and /t/-final words (RT in fair-SEAT trials minus RT in seed-SEAT trials) for participants exposed to critical words (experimental group) or replacement words (control group). Error bars represent standard error of the mean.

Discussion

The results revealed no significant effects of exposure group for either experiment, suggesting that participants were either not learning the accented /d/ category or were not able to generalize to the items presented during test. Originally, Experiment 2a and 2b were designed to implement situations where a) listeners heard speech produced with a native English-like accent and then were asked to generalize to a typical Mandarin-accented speech (Talker 2 to Talker 1), and b) listeners heard speech sounds produced with typical Mandarin-accented acoustic values and then to generalize to English-accented speech (Talker 1 to Talker 2). Since only the order was switched, acoustic differences across exposure and test were kept the same. Therefore, if prior perceptual experience influences generalization, we should have observed a significant effect of exposure group, in Experiment 2b but not 2a. However, our hypothesis was not supported.

Nevertheless, as stated in the Discussion of Experiment 1, we noted one reason that we may have failed to observe learning effects overall and to replicate the original study of Xie & Myers was that the acoustic alignment between the monosyllabic words and multisyllabic words may have naturally differed across talkers across studies. If this was the case, Experiment 2a would have created a situation in which learning might occur because the 80 ms extension on the critical exposure words made the acoustic distribution align with the critical test word distribution. In contrast, Experiment 2b would then have provided the largest acoustic differences over which listeners needed to generalize. Although we failed to find expected differences across experiments, the numerical trends are consistent with generalization of learning in Experiment 2a but not Experiment 2b. Of course, since our analyses revealed no significant differential priming effect as a finding of exposure condition, additional work needs to be done.

General Discussion

This study was designed to investigate whether prior perceptual experience would facilitate cross-talker generalization. In Experiment 1, we attempted to replicate Experiment 1 in Xie, Theodore and Myers (2017). However, we did not observe any learning or generalization in our task. The failure to replicate abolished the foundation for our original hypothesis. Since the talker we have in the current study did not produce monosyllabic words (during exposure) and multisyllabic words (at test) within a similar acoustic range, we hypothesized that the significant difference between the exposure and test words may have caused the failure.

Experiment 2 examined differences in vowel duration distributions from exposure to test, with those distributions representing more or less Mandarin-accented tokens. One consequence of the vowel duration manipulation was that in Experiment 2a, the acoustic differences between exposure and test were minimized while in Experiment 2b, those differences were maximized. Since Experiment 2a presented the smallest acoustic differences between critical exposure and test words and Experiment 2b presented the largest, we predicted that learning and generalization would occur in 2a but not in 2b. However, in Experiment 2, we again did not observe any significant effects of exposure condition, in contrast to previous work. Therefore, the data did not support any of our hypotheses regardless of whether it was the original or the updated. Nevertheless, comparing the two sets of hypotheses, the numerical patterns may have been more consistent with our new hypothesis.

One reason that we did not find any significant results may have been due to the fact that we collected the data online. Compared to the implementation of the same paradigm in the lab (Xie, Theodore & Myers, 2017; Xie & Myers, 2017), the online experiments appeared to produce higher variability in the dataset. In all three experiments that we conducted, we had more subjects than Xie and Myers had in their study, but the standard deviations (see Table 1 and 2) for the mean RTs in each prime type and exposure group were numerically higher than they were in previous studies. This high variability might lead to an issue of power. Our future plan is to re-run this data collection at more controlled environment such that we can avoid all the downsides brought up by online data collection. Another possible reason is that the talker from whom our stimuli were recorded was relatively difficult to adapt to. According to the previous studies that investigate cross-talker adaptation (Xie & Myer, 2017; Xie, Liu & Jaeger, 2021), speech adaptation would strictly follow a talker-specific manner regardless of which paradigm is adopted. Therefore, we may conclude that the null effect in the current study is due to the particular hard-to-learn production from the talker’s speech.

References

Alexander, J. E., & Nygaard, L. C. (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *145*(6), 3382-3398.

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior research methods*, *53*, 1407-1425.

Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.10, retrieved 3 May 2023 from http://www.praat.org/

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 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.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 521.

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.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias?. *Perception & psychophysics*, *70*, 604-618.

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.

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *journal of phonetics*, *32*(1), 111-140.

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of phonetics*, *35*(3), 421-438.

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of phonetics*, *35*(3), 421-438.

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

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, *116*(4), 752.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(5), 1166.

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

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

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of phonetics*, *27*(4), 359-384.

Kapnoula, E. C., & Samuel, A. G. (2019). Voices in the mental lexicon: Words carry indexical information that can affect access to their meaning. *Journal of Memory and Language*, *107*, 111-127.

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

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *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), 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.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the acoustical society of America*, *29*(1), 98-104.

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.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental science*, *12*(3), 369-378.

Melguy, Y. V., & Johnson, K. (2022). Perceptual adaptation to a novel accent: Phonetic category expansion or category shift?. *The Journal of the Acoustical Society of America*, *152*(4), 2090-2104.

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. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204-238.

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

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.

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

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*(1-2), 227-256.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306-3316.

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.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273-13278.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*, 2064-2072.

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, *97*, 30-46.

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*(11), e22.

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

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