The Relationship Between Phonemic Category Boundary Changes and Perceptual Adjustments to Natural Accents

Yi Zheng [1] and Arthur G. Samuel [1,2,3]

1. Department of Psychology, Stony Brook University
2. Basque Center on Cognition, Brain, and Language
3. Ikerbasque, Basque Foundation for Science

Contact Information:

Yi Zheng
Dept of Psychology
Stony Brook University
Stony Brook, NY 11794-2500
Email: yizheng.psychology@gmail.com

Abstract

People often experience difficulties when they first hear a novel accent. Prior research has shown that relatively fast natural accent accommodation can occur. However, there has been little investigation of the underlying perceptual mechanism that drives the learning. The current study examines whether phonemic boundary changes play a central role in natural accent accommodation. Two well-established boundary shifting phenomena were used here -- recalibration and selective adaptation -- to index the flexibility of phonemic category boundaries. Natural accent accommodation was measured with a task in which listeners heard accented words and nonwords before and after listening to English sentences produced by one of two native Mandarin Chinese speakers with moderate accents. In two experiments, participants completed recalibration, selective adaptation, and natural accent accommodation tasks focusing on a consonant contrast that is difficult for native Chinese speakers to produce. We found that (1) On the accent accommodation task, participants showed an increased endorsement of accented/mispronounced words after exposure to a speaker's accented speech, indicating a potential relaxation of criteria in the word recognition process; (2) There was no strong link between recalibrating phonemic boundaries and natural accent accommodation; (3) There was no significant correlation between recalibration and selective adaptation. These results suggest that recalibration of phonemic boundaries does not play a central role in natural accent accommodation. Instead, there is some evidence suggesting that natural accent accommodation involves a relaxation of phonemic categorization criteria.

Keywords: speech; accent; recalibration; selective adaptation; phonemic category

The Relationship Between Phonemic Category Boundary Changes and Perceptual Adjustments to Natural Accents

Picture yourself sitting in a room while listening to a speaker with a strong accent. The speech is hardly intelligible at the beginning, but after a while, you can understand most of the content. In the modern world, such situations are increasingly common, especially in a multicultural environment. People often experience difficulties when they first hear a novel accent, either regional or foreign. However, the human perceptual system can adjust to new accents, leading to improved performance in accented speech perception and comprehension (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004).

Although there are many empirical demonstrations of perceptual adjustments to accented speech, there is no consensus on the mechanism(s) that may produce the observed improvement. In the current study, we test one increasingly popular suggested mechanism – listeners recalibrate phonemic categories, and adjust the boundaries between them, as a function of the non-standard pronunciations that they hear. We begin by briefly summarizing some of the demonstrations of adjustments to accented speech, followed by a description of mechanisms that have been proposed to produce these adjustments. We then outline the approach that we use to test whether phonemic recalibration is a likely mechanism for accent accommodation. Our approach is to obtain measures of recalibration-based boundary shifts, adaptation-based boundary shifts, and accented-speech accommodation for a large number of listeners, and to see whether the accent accommodation scores correlate with the boundary shifts: If recalibration is a major mechanism for accent accommodation, the two should covary.

In developing our procedures, we needed to make a fundamental decision about the amount of exposure to an accent that is needed for listeners to make adjustments. We will thus briefly review some prior tests of perceptual adjustment to accented speech, focusing on the amount of exposure used in each study.

Perceptual Adjustment to Accented Speech

Researchers have used a number of methods to study perceptual adjustments to accented speech. In a widely cited paper, Clarke and Garrett (2004) used a cross-modal word matching paradigm in which participants heard sentences recorded from native or foreign-accented speakers. The final word of each sentence was the target word, and each sentence was followed by a visual probe word that could differ from the target word by one phoneme. Participants judged whether the visual probe matched the target word. Clarke and Garrett found that the learning process (indexed by a decreasing difference in reaction time between the accented speech and native speech stimuli) occurred within one minute of exposure, in some cases requiring fewer than four sentences.

Floccia, Goslin, Girard, and Konopczynski (2006) presented participants with sentences produced with the home accent (i.e., familiar accent), an unfamiliar regional accent, or a foreign accent. The listeners made lexical decisions about the last item of each sentence, and significant costs for accented speech remained after 32 sentences. This finding is inconsistent with the rapid adaptation found by Clarke and Garrett, possibly because Floccia et al. used multiple accented speakers, while Clarke and Garrett had used only one accented talker.

Studies that have used transcription accuracy as the measure of understanding accented speech have demonstrated relatively fast adjustment, though not as fast as Clarke and Garrett's (2004) results. Bradlow and Bent (2008) asked participants to transcribe Chinese-accented sentences and found that adjustment was completed within the 64-sentence session. Using similar procedures, Gordon-Salant, Yeni-Komshian, Fitzgibbons, and Schurman (2010) found that perceptual adjustments to Spanish-accented speech were made within the first 40 sentences.

Witteman, Weber, and McQueen (2013) found that the speed of accent accommodation was modulated by the listeners' experience with an accent as well as the speakers' accent strength. Listeners with limited experience with a German accent did not adapt to a strong German accent during a cross-modal priming task within the first eight minutes, contrary to Clarke and Garrett's (2004) finding of very rapid adjustment.

As this brief review indicates, there is a fair amount of variation across studies in the amount of exposure to accented speech that is needed for substantial accommodation to occur. This variation presumably depends on differences in the measures used, the nature of the accented speech, and the knowledge of the listeners being tested. Looking across the studies, it seems reasonable to expect substantial accent accommodation to occur if listeners are given at least 50-60 accented sentences to listen to. Based on this review, we designed an exposure phase with about a hundred accented sentences.

Mechanisms of Accent Adjustments

Despite the extensive findings that listeners adjust to accents relatively quickly, it is still unclear what underlying perceptual mechanism produces the observed accent accommodation. There are at least three potential mechanisms: (1) remapping one native sound onto another, when an accent produces such a change, (2) recalibration of phonemic category boundaries, and (3) phonemic criteria relaxation. Sumner (2011) examined the first case, looking at the "bad map" that occurs when a native French speaker produces English /p/; the French voice onset time for /p/ maps directly onto the voice onset time for English /b/. In the current study, although we will consider the “bad map” case, our primary focus is on less extreme accent-induced changes, ones that may yield more ambiguous sounds.

Recalibration studies have examined these cases that involve phonetic ambiguity, using stimuli in which experimenters create the ambiguous sounds (note that some authors use “retuning”, and some use “perceptual learning” rather than “recalibration”). Such stimuli are used to investigate adjustments to phonetic variation with a specific focus, one not directly tied to accent-based variation. For example, in the seminal recalibration study, Norris, McQueen, & Cutler (2003) examined how listeners adjusted the perceptual boundary between /s/ and /f/ as a consequence of hearing sounds midway between /s/ and /f/ (with this case not being tied to any particular natural accent). Studies associated with the criteria relaxation idea are more diverse, but share the view that listeners loosen their acceptance criteria (at multiple levels, e.g., phonetic and syntactic) in accommodating to unfamiliar accents. The empirical focus of the current study is a test of whether recalibration is a major mechanism supporting accent accommodation; our secondary focus is on a possible role for criteria relaxation. Therefore, in the following two sections, we review the most relevant studies in the recalibration and the criteria relaxation literatures. It is of course possible that both mechanisms might play a role. As noted, our primary focus is testing whether we can find evidence for recalibration playing a substantial role.

Recalibration of Phonemic Category Boundaries

Norris, McQueen, and Cutler (2003) developed a lexically-driven recalibration paradigm to investigate people’s perception of artificial idiosyncratic pronunciations (see Bertelson, Vroomen, & de Gelder, 2003 (and many succeeding studies) for a similar effect based on lip movement information rather than lexical context). In their recalibration experiments, participants performed a lexical decision task followed by a phonemic categorization task. On the lexical decision task, participants heard items with some phonemes modified (e.g., a naturally produced phoneme /s/ replaced by an ambiguous sound midway between /s/ and /f/) and decided if the items were words or non-words. Critically, for some listeners, the ambiguous sound was always presented in words in which there should be an /s/, while for other listeners the ambiguous sound always replaced /f/. On the phonemic categorization task, participants identified ambiguous sounds (ranging from /s/ to /f/) as one category (/s/) or the other (/f/). Norris et al. (2003) found that the phonemic category boundaries were shifted by prior exposure to the lexically biased contexts. If participants heard ambiguous /f/-/s/ sounds in words with lexical context consistent with /f/, then /f/ responses increased after recalibration; a comparable effect was found for the /s/ context. Norris et al. (2003) suggested that such modulation of existing phonemic category boundaries can be relevant to the fast learning of unfamiliar speech, including accented speech.

Using this recalibration methodology, several studies have shown that the category shifts are usually talker-specific (Eisner & McQueen, 2005; Kraljic, Brennan, & Samuel, 2008). However, Kraljic and Samuel (2006) provided evidence that the category recalibration generalizes under certain conditions (also see Xie, Earle, & Myers, 2018 for related findings, though in a somewhat different paradigm). Kraljic and Samuel exposed participants to idiosyncratic pronunciations (sounds ambiguous between /d/ and /t/) in /d/- and /t/-biased contexts during the lexical decision task, and found that the phonemic category shifted not only on a /d/-/t/ test continuum but also on a /b/-/p/ continuum. This effect was observed even when the exposure and test stimuli were presented in very different voices. Kraljic and Samuel (2007) demonstrated that the perceptual system was able to maintain multiple representations for some phonemes (fricatives) but not others (stop consonants). Eisner and McQueen (2006) found the effect 12 hours after the original lexical exposure, with or without sleep (and thus, possible consolidation), raising the possibility that the learning lasts long enough to be relevant to accent adjustment. Xie, Earle, and Myers (2018) also demonstrated stable learning after 12 hours, though their experiments did not use the ambiguous-token procedure that is the norm for recalibration studies.

The many studies that followed Norris et al. (2003) leave no doubt that lexically driven recalibration adjusts phonemic categories in response to ambiguous pronunciations (e.g., Eisner & McQueen, 2005, 2006; Eisner et al., 2013; Kraljic et al., 2008; Kraljic & Samuel, 2005, 2006, 2007). In a number of these studies, the authors suggested that the observed recalibration could be a mechanism for listeners' adjustments to accented speech. For example, Eisner et al. (2013) stated that "native English listeners adjusted readily to word-final devoicing of stops, both in Dutch-accented and in native-accented English. The learning mechanism driving this adjustment appears to generalize relatively broadly in foreign-accented speech after only limited exposure" (p. 8). Kraljic and Samuel (2006) also associated the recalibration of stop consonants with natural accent accommodation by asking "whether … listeners learn a particular speaker's "accent"." (p.262).

Reinisch and Holt (2014) explicitly tried to link the findings of recalibration studies to naturally accented speech processing. They embedded artificially manipulated ambiguous fricatives in natural speech and found that listeners adjusted their phonemic categories within the context of a global foreign accent. In other words, the typical recalibration effect still holds, even when the contextual surroundings are clearly accented. The authors said "by presenting an artificially manipulated phoneme contrast in an unfamiliar global foreign accent, we have demonstrated across three experiments that lexically guided phonetic category recalibration is observable in the context of a global foreign accent and, thus, could play a role in adaptation to naturally occurring foreign accents" (p.20). We note, however, that these results demonstrate that the presence of a noticeable accent does not block recalibration; they do not demonstrate that phonemic category shifting plays a substantial role in the accommodation of naturally accented speech any more than other recalibration studies do.

In addition to recalibration, another phenomenon -- selective adaptation -- also produces shifts in listeners' phoneme boundaries. Selective adaptation occurs after hearing unambiguous speech sounds repeatedly, with listeners reducing their report of similar-sounding phonemes (Eimas & Corbit, 1973; Samuel, 1986). For example, after hearing a /d/ sound repeatedly, people's perception of /d/ is reduced, measured by their identification of, e.g., stimuli making up a /d/- /t/ continuum. Kleinschmidt and Jaeger (2015) have argued that recalibration and selective adaptation are different manifestations of the same adjustment principles. It is unknown yet whether or not these category adjustment principles account for a significant portion of accent adjustment.

Criteria Relaxation

An alternative mechanism that has been suggested for accent accommodation is the relaxation of identification criteria in foreign-accented speech perception. The "relaxation of criteria" idea is that listeners increase their tolerance of non-standard speech patterns from non-native speakers by reducing the threshold for accepting a stimulus as being a particular phoneme, word, or syntactic structure; they become more flexible about the sounds or structures that constitute speech. For example, Hanulíková and colleagues (2012) found that gender disagreements (i.e., inconsistency between the definite determiner and the noun in Dutch sentences) elicited a P600 (an ERP component associated with syntactic repair processes) if the error was produced by a native speaker, but no such effect was found if it was from a foreign-accented speaker. These results are in line with Reinisch and Weber (2012)'s report that listeners rapidly adjust to suprasegmental stress errors in foreign accents. Participants listening to such accented speech successfully adjusted to stress errors. Schmale, Cristia, and Seidl (2012) explicitly embraced this idea in the context of their argument that, unlike the phonemic boundary shifts caused by recalibration, natural accents typically involve variations along multiple dimensions, and thus a "general expansion" strategy can be useful in such circumstances.

The Current Study

In our brief review of the recalibration literature, we noted that this phenomenon is consistently described as a likely mechanism for perceptual adjustment to accented speech (the argument is not that this is the only type of adjustment, as other adjustments, e.g., to rhythmic differences, presumably occur). This claim has clear face validity: When accented speech does not produce a full change of one phoneme into another (the "bad map" case described by Sumner, 2011), it produces ambiguous sounds that could be quite similar to the experimenter-produced ambiguous sounds in recalibration experiments. Despite this apparent plausibility, there is actually no direct evidence to support the suggestion. Only one recalibration study explicitly set out to test the connection between recalibration and accent accommodation (Reinisch & Holt, 2014), but as we noted, this study only shows that the recalibration effect is not blocked when the carrier speech is accented – this is quite different than showing that recalibration is the mechanism driving accent accommodation. A set of studies by Xie, Myers, and their colleagues (e.g., Xie & Myers, 2017; Xie, Earle, & Myers, 2018; Xie, Theodore, & Myers, 2017) is more directly tied to accent accommodation, but does not use experimenter-generated ambiguous sounds in a standard recalibration study.

Given the boom in recalibration studies, including those from our laboratory, and given that the seminal lexical recalibration study (Norris et al., 2003) was reported over 15 years ago, we believe it is time for researchers in this area to explicitly test whether recalibration really plays a major role in perceptual adjustments to accented speech. Here, we report such an explicit test. We do not claim that our test is definitive (any single study is rarely, if ever, definitive), but we believe it is a very good initial attempt to test the claim that we and others have made. Our intent is to stimulate others to conduct their own explicit tests so that the resulting body of research will either confirm the standard claim, or refute it.

The logic of our test is built on the idea that individuals may vary in the extent to which they adjust to non-standard speech sounds. Assume that there are two speech sounds, X and Y, and that in accented speech X is produced as some ambiguous version, somewhere between the native X and Y sounds (we will call the accented version of X '***x_y***'). Further assume that some listeners can adapt to relatively extreme versions of ***x_y***, ones that are relatively far from X, whereas other listeners can only adapt to versions of ***x_y*** that are closer to the original X. If the mechanism that supports adaptation to ***x_y*** is recalibration (or is the same mechanism that produces recalibration, if we take the observed recalibration in experiments to be a measure of an underlying mechanism's effect), then we should see similar individual differences on a recalibration test that uses experimenter-generated ambiguous tokens between X and Y: Listeners who show large accent accommodation for ***x_y*** should show large recalibration shifts for ***x_y***, and listeners who show less accent accommodation should show smaller recalibration shifts. The idea is that if there is in fact one underlying mechanism (which has been the claim in recalibration studies), that mechanism should operate in the same way within an individual, whether the measure is accent accommodation or recalibration. If instead recalibration is not directly related to natural accent accommodation (i.e., there are different underlying mechanisms, each of which can affect phonetic categorization measures), then there should be no systematic relationship between the two cases across listeners.

As the preceding description should make clear, the core test in the current study is correlational. There are some additional aspects that we will describe, but fundamentally the question we ask is whether a measure of accent accommodation to a particular non-native production correlates with a measure of recalibration for the same kind of sound change when the change is implemented in the standard recalibration paradigm. Finding such a correlation would provide evidence for the claim that has frequently been made regarding the function of recalibration: The mechanism driving shifts in recalibration studies is the same mechanism that plays a significant role in accent accommodation. If we fail to find such a correlation, then assuming that the test meets certain criteria, the onus will be on researchers to produce positive evidence that recalibration effects are in fact related to accent accommodation.

There are three criteria that would need to be met for a null correlation to be evidence against that claim. First, as with any null effect, a power analysis would need to indicate that the test had a sufficient sample size. Second, there would need to be evidence that there is sufficient between-subject variance to allow a correlation to emerge; if all of the scores are very tightly clustered on one measure and/or the other, correlations cannot be found. Finally, the effects (recalibration, accent accommodation) need to be stable enough within an individual so that a correlation can succeed; if an individual produces a large recalibration effect one day, and a small one on a different day (for the same test), it will not be possible to correlate that score with another score. Because of the nature of the recalibration effect (see the discussion below of the “inoculation” issue), it is actually very difficult to get a within-subject test-retest measure, making it difficult to assess the stability question. One prior study (Saltzman & Myers, 2018) did collect such test-retest scores, and the stability was not strong. Of course, if recalibration is in fact the basis of accent accommodation (or, if a common underlying mechanism drives both), it would need to be stable enough to support such accommodation in the first place – a recalibration that comes and goes would not allow a listener to maintain improved understanding of accented speech.

We have described the logic of our test in terms of the recalibration effect's correlation with a measure of accent accommodation, but we will also measure the size of selective adaptation shifts for each participant as well. As we noted above, in Kleinschmidt and Jaeger's (2015) influential paper, adaptation and recalibration are treated as two consequences of the same computational mechanism. Thus, having both measures will provide two estimates of boundary shifts for each listener. In addition, by collecting both measures for each participant, we can test whether there is any correlation between these two effects, as might be expected if they are in fact driven by the same mechanism.

In the current study, the recalibration and selective adaptation paradigms are applied to an English consonant contrast that was chosen because it is difficult for Mandarin Chinese speakers (Flege, Bohn, & Jang, 1997; Rogers & Dalby, 2005; hereafter, "Chinese" will be used to refer to Mandarin Chinese). Specifically, the /θ/-/s/ contrast (the initial consonants in "think" and "sink", respectively) was selected because Chinese has one side of the contrast (i.e., /s/) but not the other (/θ/), which causes difficulties for native Chinese speakers in producing the non-native phonemes. In fact, while /θ/ exists in English, many other languages do not have this phoneme, causing difficulties for many L2 English learners. For instance, the /θ/-/s/ contrast is difficult for German and Dutch speakers (Brannen, 2002; Cutler, Weber, Smits, & Cooper, 2004; Hancin-

Bhatt, 1994; Hanulíková, & Weber, 2012). The current study used this contrast to measure how much listeners accommodated to naturally-produced Chinese-accented English, focusing on the amount of adjustment in understanding the chosen phonemes as a function of exposure to the accented speech. Two non-native speakers were chosen who met our criterion of having productions of /θ/ that deviate from the standard American English pronunciation (the acoustic differences are discussed below).

In order to conduct the desired test, careful construction of stimuli and tasks was essential for the recalibration, selective adaptation, and natural accent accommodation tasks. Because our approach relies on correlation tests to assess the relationship among natural accent accommodation, recalibration, and selective adaptation, the analysis requires within-subject measurement of the contrast for all three tasks. For the recalibration and selective adaptation tasks, this means that measures from both sides of the contrast (e.g., both /θ/ and /s/) within each subject were needed. However, because recalibration is very delicate -- it can be blocked by previous experience with hearing "good" tokens in the speaker's voice -- measuring one side of a contrast may contaminate measurement of the effect on the other side. Kraljic et al. (2008) called this disruption an “inoculation effect”, as an initial exposure to good category members can prevent recalibration by ambiguous tokens. Extensive pilot testing was undertaken in order to identify procedures that were likely to produce reliable measures for each task within individual participants, despite potential inoculation effects.

Pilot Tests

The purpose of the recalibration effect piloting was to find conditions that would provide a within-subject measurement of both sides of a contrast. The factors expected to influence the results included the speaker’s voice, the amount of exposure to a given set of stimuli, and the amount of time between exposure to one set of stimuli and exposure to that set or another set afterwards. A series of pilot tests led to the final design used for the recalibration task: Participants were tested on both sides of the /θ/-/s/ contrast, (a) with both sides of the contrast presented in the same (male) voice; (b) with a two-week spacing between the two sessions; and (c) with relatively few presentations of the stimuli during each test. With these procedures, reliable and stable shifts were obtained.

Selective adaptation has proven to be very robust across many methodological manipulations. In piloting the selective adaptation task, a 7-step /θ/-/s/ continuum in a female voice was tested. We chose a female voice in order to avoid any interference from this task on the recalibration task (which used a male voice); recall that recalibration effects are generally speaker-specific. Consistent with the robust adaptation effects in the literature, the pilot testing produced large shifts in the expected direction: repeated presentation of /θ/ reduced report of /θ/, and repeated presentation of /s/ reduced report of /s/.

The goal of the accent accommodation piloting was to find a reliable measure of adjustment to naturally-produced Chinese accented speech, with a reasonable training and testing time. We adopted a “pretest-training-posttest” approach, looking for training-induced improvement from the pretest to the posttest. Several tests were run using minimal pairs (e.g., recognizing “sink”, or recognizing “think”) in the pretest and posttests, but no significant change was observed from pretest to posttest. This result led us to a different approach for the pretest and posttest: Listeners were presented with words and non-words, in a lexical decision task. The stimuli included *critical words* that contained the critical phonemes (e.g., /θ/ in “withdraw”). *Critical non-words* were created by replacing the critical phonemes with the other side of the contrast (e.g., “wisdraw” with /s/ substituted for /θ/). These stimuli did not contain any lexical minimal pairs. In addition to critical words and non-words, *filler words* (e.g., “catch”) and *filler non-words* (e.g., “gatch”) were also created. The fillers did not contain the critical sounds being tested. During the training period, listeners heard 96 simple Chinese-accented English sentences, produced by the same native Chinese speakers who produced the items for the pretest and the posttest.

After the training phase, listeners accepted more items as words than they had on the pretest. This change entails an improvement from pretest to posttest for the critical words, but a decrease in accuracy for the critical non-words (since the correct answer should have been “non-word”); the filler items produced a similar pattern with smaller magnitude. In a series of pilot tests, adjustments were made to sort items into lists for the pretest and posttest that were as well matched as possible. The sorting was done in order to optimize two average values, a “difficulty score”, and a “difference score”. For each pilot test, a “difficulty score” (the average accuracy across participants) was obtained for each item, along with a “difference score” (the average accuracy of an item when it was presented in the posttest minus its accuracy in the pretest). After each iteration of pilot testing, the item distribution was modified for the lists used in the pretest and posttest, balancing the overall difficulty scores, difference scores, word durations, and word frequency (for both critical and filler words and non-words). The final result was two lists (List A and List B) that were to be counterbalanced in the main test.

Finally, we did a comprehensive set of acoustic analyses to determine that we had two non-native speakers whose /θ/ - /s/ productions differed systematically from native speakers. Six parameters were chosen: duration, frequency, and spectral moments (center of gravity, skewness, kurtosis, and standard deviations), based on the prior literature (e.g., Jongman et al., 2000; Kent et a., 1992). Density plots for six native speakers were created to aid visualization. These plots made it clear that three of the consonant parameters (duration, frequency, center of gravity) reliably differentiated native from non-native productions. Appendix 1 provides the density plots, based on speakers' production of the items in Appendix 3. Two Chinese speakers (of six whose productions were analyzed) were selected because their productions were clearly non-native. As can be seen in the plots, both speakers produced /θ/ in a nonnative way, with the modal values being relatively close to those for /s/. For both speakers, the distribution of tokens was noticeably wider than native (i.e., the tokens were much more variable), with this being especially true for Speaker 2.

Main Study

In the following two experiments, we investigate the relationship between phonemic category boundary change and natural accent accommodation for the /θ/-/s/ contrast. In keeping with current concerns about the replicability of experiments, the main study includes two experiments that only differ in the non-native speakers who produced the to-be-accommodated speech. For clarity of exposition, we will call the data collection based on one speaker "Experiment 1", and the data collection based on the other speaker "Experiment 2".

Method

Participants

Undergraduate students at Stony Brook University were recruited for the current study. All were native English speakers, 18 years of age or older, with self-reported normal hearing and vision. Participants received research credits and payment for their participation. This research was approved by the Stony Brook University Institutional Review Board (Study Title: "Bilingual Language Use", Stony Protocol #119040).

The number of participants to test in each experiment was determined based on sample estimation for correlation studies. With α = .05, an assumed moderate correlation size (r = .40), and a desired power of 0.8, a sample size of 47 is needed. To assure this power level, data from 52 usable participants were collected in each experiment. Thus, across the two Chinese speakers, we tested a total of 104 usable participants. Up to three participants were tested at the same time in a sound-attenuated booth.

Design

Each participant took part in the natural accent accommodation, recalibration, and selective adaptation tasks. All participants were tested with the same recalibration and selective adaptation tasks, but on the accent accommodation task, half of the participants heard one Chinese speaker (referred to as Chinese Speaker 1) [Experiment 1] and the other half heard the other (Chinese Speaker 2) [Experiment 2].

The three tasks were spread over three sessions, with exactly one week between successive sessions. Table 1 shows how the tasks were distributed in each experiment. In Session 1, each participant did one side of the recalibration task (e.g., the /s/ condition, in which all critical /s/ items, such as *legacy*, had /s/ replaced with a sound midway between /θ/ and /s/), and the same side of the selective adaptation task. The opposite side of the contrast, for both recalibration and selective adaptation, was tested two weeks later, in Session 3.

The recalibration tests were presented in a male voice and the selective adaptation tests were presented in a female voice, as prior research provided evidence that recalibration is largely talker-specific (Eisner & McQueen, 2005; Kraljic et al., 2008). Having two different speakers whose voices are quite different was intended to maintain the independence of the recalibration and the selective adaptation tests. The two recalibration sessions were separated by two weeks to minimize possible contamination effects of the first recalibration test on the second; earlier work showed that the recalibration blocking effect of prior exposure to unambiguous items still operated after one week (Zhang & Samuel, 2014). The design provided both sides of each recalibration case for each individual participant, and also provided a within-subject score for selective adaptation for the critical contrast. Note that each listener was exposed to three different speakers, a male native English speaker (recalibration), a female native English speaker (selective adaptation), and a female native Mandarin speaker (natural accent accommodation). The Mandarin speaker (Speaker 1 or Speaker 2) was not used for the recalibration or adaptation stimuli for two reasons. First, the Mandarin speakers were chosen because their productions of /θ/ were quite non-native, making it impossible to generate a /θ/ - /s/ continuum. Second, exposure to the Mandarin speaker in Session 2 would presumably block recalibration in Session 3 if the same voice was used. There was no need to match voices for the correlational test because we are investigating whether a common mechanism drives the different effects, and such a mechanism would not be voice-specific (as opposed to recalibration itself, which in most cases does not transfer across voices).

In Session 2 (one week after Session 1), participants did the natural accent accommodation task. As described in the summary of the pilot tests, two balanced lists were created for use in the pretest and the posttest. Half of the participants heard List A in the pretest and List B in the posttest (in Table 1, “pretest List A, posttest List B”), and the other half heard the two lists in the reverse order. These two counterbalancing orders were crossed with the order of testing the two sides of the contrast in Sessions 1 and 3 (/θ/ versus /s/), yielding four groups of participants. This design provides the basis for a by-subject correlation with accent adjustment:

For each participant, there was a measure of recalibration, selective adaptation, and natural accent accommodation for the critical contrast. Table 1 shows the design of the two experiments.

Table 1. The experimental design used in each Experiment

Group	Session 1	Session 2	Session 3
G1	Recalibration-/s/ Selective Adaptation-/s/	Natural Accent Accommodation: (pretest List A, posttest List B)	Recalibration-/θ/ Selective Adaptation-/θ/
G2	Recalibration-/s/ Selective Adaptation-/s/	Natural Accent Accommodation: (pretest List B, posttest List A)	Recalibration-/θ/ Selective Adaptation-/θ/
G3	Recalibration-/θ/ Selective Adaptation-/θ/	Natural Accent Accommodation: (pretest List A, posttest List B)	Recalibration-/s/ Selective Adaptation-/s/
G4	Recalibration-/θ/ Selective Adaptation-/θ/	Natural Accent Accommodation: (pretest List B, posttest List A)	Recalibration-/s/ Selective Adaptation-/s/

Recalibration Task Materials

For the /θ/-/s/ contrast, 16 English words were selected that contain /s/ (e.g., /s/ in *legacy*), and another 16 English words were selected that contain /θ/ (e.g., /θ/ in *anything*). Words ranged from 2-5 syllables, with the critical phonemes occurring in a late position of each word, in order to ensure lexical activation of the critical phonemes. Words with one critical phoneme (e.g., /s/) did not contain the phoneme on the other side of the contrast (e.g., /θ/). All critical consonants were surrounded by vowels. The two lists of critical words were matched in average syllable length (/θ/: 3.63, and /s/: 3.50) and median word frequency (per million words: /θ/: 1.35, and /s/: 3.23) (retrieved from http://subtlexus.lexique.org/moteur2/).

All 32 critical words were recorded by a male native American English speaker. Each word was produced in two ways: a standard version, and a non-standard version in which the critical phoneme was replaced by the contrasting phoneme. For example, *legacy* with the critical sound /s/ was produced as *legacy* and as *legathy*, with the latter case including the critical /θ/ sound. The speaker was instructed to produce the non-standard version naturally, and to keep everything consistent except for the critical phoneme.

A custom-designed C++ program that our lab has used numerous times for mixing voiceless fricatives was used to merge the two versions of each word, generating a set of ambiguous tokens changing from one version (*legacy*) to the other (*legathy*). The program computes a weighted average of two waveforms, point by point (the longer of the two waveforms is trimmed from the middle, to match the duration of the shorter one). Specifically, the critical phonemes in the two versions were located, and were mixed using weighted averages, from 95% /s/ and 5% /θ/ to 5% /s/ and 95% /θ/, in 5% increments. This procedure produced a 19-step fricative continuum. Each step of the continuum was then inserted back into the original two word frames (e.g., *legacy* and *legathy*). Five native English speakers were asked to listen to all of the tokens and independently selected the most ambiguous token from the better frame for each critical word (i.e., the token in which the critical sound was perceived to be halfway between /θ/ and /s/). The final selection was made by taking the central tendency of the responses from the five listeners. The selected ambiguous mixtures for each critical item can be found in Appendix 2.

Each recalibration experiment contained 16 critical (ambiguous) words, 16 critical (unambiguous) words with the "opposite" sound, 48 filler words and 80 filler non-words. Therefore, in addition to the critical words, 48 filler words and 80 non-words were selected that did not contain either of the critical phonemes. Non-words were generated using Wuggy (http://crr.ugent.be/programs-data/wuggy), software that generates non-words from a given word seed. The filler items matched the critical words in mean syllable length and word frequency. Specifically, the filler words had a mean syllable length of 3.46 and a median word frequency of 1.12; the non-words had a mean syllable length of 3.50.

For the phonemic category identification task, the critical consonant phonemes were recorded in simple CV form. The vowel /æ/ was used so that the resulting CV syllables were both non-words. These were used to make a 7-step continuum: /θæ/-/sæ/. The construction of the /θæ/-/sæ/ continuum was done using a procedure similar to that for the critical words, which involved isolating the critical consonants, creating continuum steps with the sound mixing program, and inserting each of the continuum steps back into the two frames. A series of pilot tests was then conducted in order to find the best mixtures to use in the test continuum. Between iterations of the pilot testing, we adjusted the center point and the spacing of the mixtures. The middle step of the final 7-step continuum was identified about equally often as 's' and 'th', and the two endpoints were consistently identified as they should be.

During the phonemic category identification tasks, participants saw the labels “sa” and “tha” on a computer screen. All of the critical words, filler items, non-words, and test phonemes were recorded by the same male native speaker of American English. See Appendix 2 for a list of the stimuli that were used for the recalibration task.

Recalibration Task Procedure

Each recalibration test consisted of two phases: an exposure phase, and a phonemic categorization phase. In the exposure phase, participants performed an auditory lexical decision task. The test included 80 words (16 ambiguous-sound /θ/ or /s/, 16 unambiguous sound /θ/ or /s/, 48 filler) and 80 non-words; filler words and non-words did not have any /θ/ or /s/ sounds. All items were presented in a random order. On each trial, participants pressed one of two buttons to indicate if they heard a word or a non-word (left=word, right=non-word), with a 500 ms inter-trial interval (ITI) after the responses, and 3 sec maximum waiting time. The labels “word” and “nonword” were visible on the screen. Any missed responses were treated as wrong answers in the data analysis.

On the phonemic categorization (ID) task, 10 randomizations of a 7-step continuum were presented; the first two passes served as practice (without feedback). Participants were instructed to press the left button if an item sounded like “sa”, and the right button if it sounded like “tha”, with the labels presented on their computer screen as a reminder. The next trial began after all participants responded. If any of the participants failed to make a response within 3 sec, the program moved on to the next trial. Each participant did two recalibration tasks, one during Session 1, and one two weeks later in Session 3.

Selective Adaptation Task Materials

For the selective adaptation task, a female native American English speaker produced the critical consonant phonemes in simple CV form (/θæ/ and /sæ/). These recordings were used to make a 7-step continuum of /θæ/-/sæ/, using the procedures described for the phonemic category identification stimuli used in the recalibration task.

Selective Adaptation Task Procedure

Each selective adaptation test had two tasks: an initial baseline identification (ID) task and an adaptation test. On the first task, participants listened to 18 randomizations of the (female) 7-step continuum (with the first two passes serving as practice, without feedback). Participants identified each sound as either “sa” or “tha”. On the second task, participants made the same judgment with only one change: instead of making responses to each sound, there were periods of time when participants only listened to a sound repeating, with no response required. The adaptation test consisted of 16 cycles, with each cycle including 30 repetitions of a sound with a 500 ms inter-stimulus interval between repetitions. The repeating sound (i.e., the adaptor) was /θæ/ in one session and /sæ/ in the other. During the adaptor repeating time, no labels were presented on the computer screen, and no responses were made. The repeating-sound phase was followed by one randomization of the 7-step continuum (with labels presented on the screen). Participants were instructed to press a button to identify each sound when the labels were present; when the labels disappeared, they were to listen to the repeating sound without making any responses.

Natural Accent Accommodation Task Materials

For the accent accommodation task, 40 critical words and 40 filler words were selected (see Appendix 3). In addition, 80 non-words were created based on the 80 words. Critical non-words were made by changing the critical sound of the word (/θ/) to the substitution that occurs in a Chinese accent (/s/). For example, for the critical word “thankful”, the critical non-word “sankful” was generated. Filler non-words were created by changing one phoneme in a filler word to another phoneme that is not difficult for Chinese speakers. For example, for the filler word “catch”, the filler non-word “gatch” was generated.

These items were split into two lists: half of the items (20 critical words, 20 critical non-words, 20 filler words, and 20 filler non-words) made up List A and the other half constituted List B. If a word (e.g., “thankful”, with the critical /θ/ sound) was in List A, the corresponding non-word (e.g., “sankful”, with the /s/ substitution) was presented in List B. The two lists were well matched in syllable length, frequency, difficulty and difference scores (based on the iterations of the pilot study). The A-B list distribution was different for the two Chinese speakers, based on the different item difficulties across the two speakers in the pilot testing. For each Chinese speaker, the order of the two lists as pretest or posttest was counterbalanced across participants.

For the training part of the natural accent accommodation task (see below), 96 sentences were selected from the BKB sentence list (Bamford & Wilson, 1979). See Appendix 4.

Two female Chinese-accented speakers (selected from pilot testing and acoustic analyses) with moderate accents recorded the words, non-words, and sentences. Both speakers were instructed to produce the stimuli in a clear and natural way. Nine sentences produced by a female native English speaker were used in the practice. As described below, we gathered information from each listener about the listener’s familiarity with Chinese-accented speech.

Natural Accent Accommodation Task Procedure

The procedures developed in the final pilot test were employed in the natural accent accommodation test. As described above, the natural accent accommodation test included three phases: pretest, training, and posttest. During the pretest and posttest, each participant heard English words and non-words recorded by either Chinese Speaker 1 (Experiment 1) or Chinese Speaker 2 (Experiment 2). The order of the two sets of stimuli in the pretest and posttest was counterbalanced across participants. During the pretest and the posttest, participants listened to one item at a time, and were asked to indicate whether they heard a word or a non-word by pressing buttons on a button board (left: word, right: non-word).

During the training session between the pretest and the posttest, participants were presented with sentences spoken by the same non-native speaker who produced the materials in the pretest and posttest. All of the sentences had predictable lexical contexts (e.g., *the clown had a funny face;* the critical /s/ sound was present in the final word "face" in this example). Sentences in the training phase contained a good number of instances of the critical phoneme, providing an opportunity for the accented variation to be learned. Specifically, among the 96 sentences, the critical /θ/ sound (which tends to be more like /s/ in Chinese-accented speech) or its voiced counterpart /ð/ occurred 101 times[1].

Before the training task, participants were given a short practice version, with sentences recorded by a female native American English speaker. Participants were asked to pay close attention to the meaning of the sentences and were told that they would be tested on their understanding of the sentences. Each trial included three auditory sentences followed by one visual probe sentence. The probe sentences were used to ensure that participants understood the preceding three auditory sentences. The task was to judge whether the written sentence was a

[1] /ð/and /θ/ are the voiced and voiceless versions of the sound "th". Because they are acoustically very similar, here the number reflects the total number including both cases. Of the 101 times that "th" appeared in the training sentences, 96 were the voiced "th". The voiced "th" is typically replaced by "z" in a Chinese accent, just as the voiceless "th" is typically replaced with "s".

rewording of any of the three preceding auditory sentences (e.g., the probe "*they had a great day"* is a rewording of "*they had a lovely day"*). Participants clicked yes (left button) or no (right button). This paradigm was an adaptation of the sentence task in Zhang and Samuel (2014). The 96 sentences in the training task were divided into 32 sets of three sentences. In these 32 sets, half of the probes called for a Yes response. The probe was related to the first, second, or third sentence of each set. Note that even for mismatch responses, the probe was related to a particular sentence of each set (e.g., *"the truck broke down"* is related to *"the truck drove up the road"*, but the two sentences mismatch in meaning)

The natural accent accommodation task took 20 minutes. After that, participants finished a short language experience questionnaire (see Appendix 5).

Results

As described in the Introduction, our core question is whether natural accent accommodation correlates with the boundary shifts found for recalibration (and, perhaps, for selective adaptation). Before we report these correlations, we will describe the results for each of the three tasks (recalibration , selective adaptation, and natural accent accommodation) at the group level.

In Experiment 1, data were collected from 54 native English speaking participants. One participant was excluded because of experimenter error, and one participant was excluded due to an inability to reliably make distinctions between the two endpoints of the continuum on the baseline ID of the selective adaptation task.[2] The remaining 52 participants were evenly

[2] During Session 1 or Session 3, on the baseline ID of the selective adaptation task, if the % /s/ responses for step 1 (the least /s/-like token) was greater than 60% of step 7 (the most /s/-like token), the participant was deemed to be unwilling or unable to differentiate members of the continuum reliably (Zheng & Samuel, 2017). That is, if a participant failed to differentiate the

distributed across the four participant groups (13 participants in each of the four subgroups; see Table 1). There were 38 females and 14 males, with a mean age of 19.3 (*SD*=1.12). Data from the questionnaire indicated that participants had some familiarity with Chinese-accented English (*M*=4.7, *SD*=2.79, on a scale of 0-10, with 0=not familiar at all, 10= very familiar).

In Experiment 2, data were collected from 59 participants. Two participants failed to finish all three sessions, four participants were excluded due to experimenter error, and one participant failed to make distinctions between the two endpoints of the continuum on the baseline of the selective adaptation task. The results were analyzed based on the remaining 52 participants (34 females and 18 males, with a mean age of 20.3 (*SD*=2.63), evenly distributed in each group). The participants had some familiarity with Chinese accented English (*M*=5.8, *SD*=2.33).

Recalibration

Tables 2 and 3 show the average accuracy and RT data for each type of critical item during the exposure phase. The acceptance rate of the ambiguous items was high overall, suggesting that the ambiguous items sounded sufficiently natural. The acceptance rate was in the typical range that yields successful recalibration effects, as shown in prior work (e.g., Kraljic & Samuel, 2007; Norris et al., 2003).

Table 2. Mean Accuracy and Reaction Times for Natural and Ambiguous Critical Words in Experiment 1

	Session 1	Session 3

continuum reliably on either baseline, that participant was identified as an outlier, and thus excluded from data analysis.

	Natural		Ambiguous		Natural		Ambiguous	
	/s/	/θ/	/?s/	/?θ/	/s/	/θ/	/?s/	/?θ/
% Correct	97.4	95.7	96.4	83.7	100.0	94.0	92.5	89.2
RT (ms)	1072	1106	1114	1160	1011	977	1043	1050

Table 3. Mean Accuracy and Reaction Times for Natural and Ambiguous Critical Words in Experiment 2

	Session 1				Session 3			
	Natural		Ambiguous		Natural		Ambiguous	
	/s/	/θ/	/?s/	/?θ/	/s/	/θ/	/?s/	/?θ/
% Correct	98.1	93.3	94.7	79.1	98.3	91.1	93.0	87.5
RT (ms)	1111	1118	1166	1252	1018	1021	1048	1099

The measurement of the recalibration effect was based on the difference in phonemic category identification in the /s/ condition versus the /θ/ condition. Specifically, for each participant, the percentage of /s/ responses was calculated on the phonemic category identification (ID) tasks (during Sessions 1 and 3) for each continuum step. The first two randomizations were used to familiarize participants with the ID task and were not analyzed; the scores were calculated based on the remaining eight presentations of each continuum step. Figure 1 shows the average category identification responses on the /θ/-/s/ continuum. As expected, it shows that after exposure in the /s/ condition (i.e., hearing critical items with /s/ replaced by an ambiguous sound midway between /θ/ and /s/), people heard the test syllables as more like /s/, compared to the /θ/ condition, and vice versa.

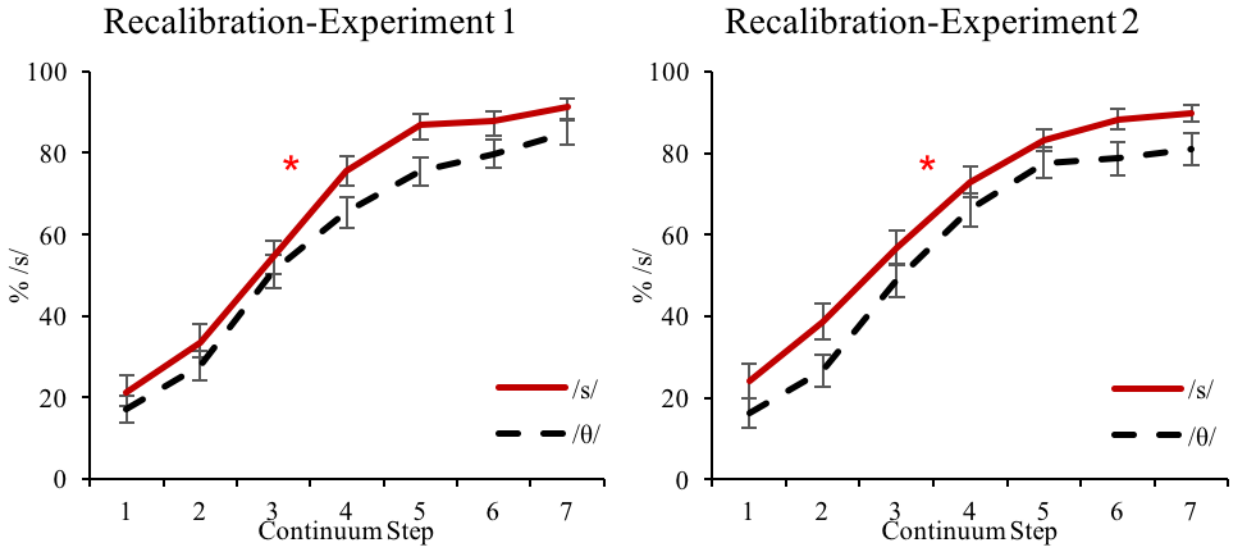


Figure 1. Percentage of /s/ responses on the 7-step /θ/-/s/ continuum as a function of recalibration condition in Experiment 1 (left panel) and Experiment 2 (right panel). Error bars represent the standard error of the mean.

For analysis purposes, the average report of /s/ on the middle three items of the seven-step continuum was used. Shifts are typically largest in the middle range of the continuum, and in fact in many studies only items in the middle of the continuum are presented for identification (e.g., Bertelson, Vroomen, & de Gelder, 2003). In our lab, we prefer to present the full range so that listeners are actually exposed to “good” tokens, but we focus the analyses on the middle items. We have examined a few different measures, and across dozens of studies, this procedure has proven to be the most sensitive one (e.g., Kraljic &Samuel, 2005, 2006, 2007; Samuel, 1986;

Zhang & Samuel, 2014). The differences of the averaged values in the /θ/ condition versus the /s/ condition were calculated for each participant. These values were used in computing correlations between tasks (see below). To statistically assess the recalibration effect itself, a three-way repeated measures ANOVA was conducted, including one within-subject factor (Condition: /θ/ vs. /s/) and two between-subject factors (Presentation Order: /s/ Session 1 and /θ/ Session 3, vs. /θ/ Session 1 and /s/ Session 3; Experiment: Experiment 1 vs. Experiment 2). The main effect of Condition was significant, *F* (1, 100) =12.27, *p* =.001, η2 = .109, reflecting significant recalibration. The interaction between Condition and Experiment was not significant, *F* (1, 100) =.24, *p*=.629, η2 = .002, indicating that, as Figure 1 suggests, the two Experiments produced similar recalibration effects. There was no difference between groups due to Presentation order, *F* (1, 100) =1.35, *p* =.248, η2 = .013, and no difference between the two Experiments, *F* (1, 100) =.048, *p* =.827, η2 < .001. The interaction between Condition and Presentation order was not significant, *F* (1, 100) =.026, *p* =.872, η2 < .001.

Further analyses were conducted to examine the between-subject recalibration effect during each of the two sessions, collapsing across the two experiments. A one-way ANOVA was conducted using Presentation Order as the independent variable, and averaged /s/ responses across the middle three steps of the continuum in Session 1 as the dependent variable, contrasting differences of the /s/ and /θ/ conditions in the first session. Similarly, another one-way ANOVA was conducted using Session 3 results as the dependent variable. Figure 2 shows the recalibration effect for Session 1 (left panel) and for Session 3 (right panel). As the figure shows, the shift was substantial in Session 1, but smaller in Session 3. A significant recalibration effect was found in Session 1: *F* (1, 102) =7.74, *p* =.006, but not in Session 3, *F* (1, 102) =.56, *p* =.454. This analysis indicates that a certain amount of “inoculation” remained, despite the procedures that were piloted extensively: a smaller amount of exposure than usual (i.e., fewer passes through the testing continuum) and a two-week separation between sessions.

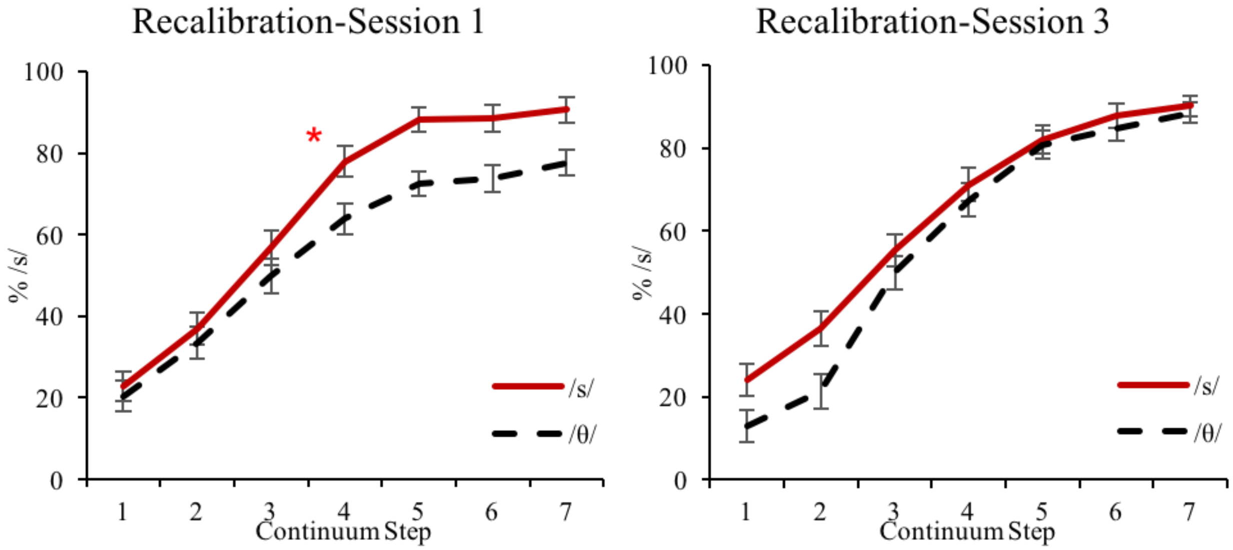


Figure 2. Percentage of /s/ responses on the 7-step /θ/-/s/ continuum as a function of recalibration condition in Session 1 (left panel) and Session 3 (right panel). Error bars represent the standard error of the mean.

Selective Adaptation

The measurement of the selective adaptation effect used a similar approach to that for recalibration. The percentage of /s/ responses was calculated for each participant on the baseline and adaptation identification tasks, in both Sessions 1 and 3, for each continuum step. The first two randomizations on the baseline tasks were used to familiarize participants with the task and were not analyzed, so the scores were calculated based on the remaining 16 repetitions of each continuum step[3].

[3] In Experiment 1, due to a computer failure, the data from three participants on the adaptation task had only 10 instead of 16 randomizations; for these three participants, scores were based on 10 observations per continuum step.

Figure 3 shows the average category identification responses on the /θ/-/s/ continuum, both before and after adaptation for the /θ/ and /s/ conditions, collapsing across the two presentation orders. The adaptation effects were extremely large, with the boundary shifting as expected: After hearing many repetitions of /s/, responses of /s/ were reduced, with a comparable effect of /θ/ repetition.

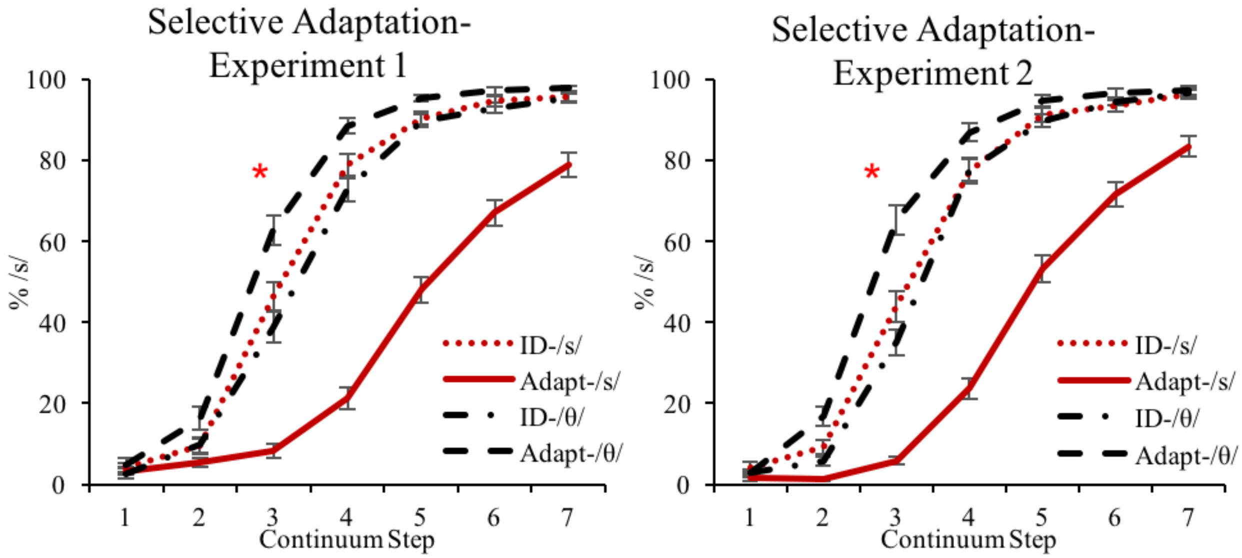


Figure 3. Percentage of /s/ responses for each step on the 7-step /θ/-/s/ continuum on the baseline ID and adaptation tasks as a function of adaptation condition in Experiment 1 (left panel) and Experiment 2 (right panel). Error bars represent the standard error of the mean.

For analysis purposes, the average report of /s/ on the middle three items of the seven-step continuum was used. The difference in baseline identification versus identification after adaptation in the /s/ condition indexed the phonemic category boundary change during one session, with the same measure used in the /θ/ condition for the other session. The difference of these two values (i.e., generally a positive number minus a negative value) was used as the measure of the selective adaptation effect for each participant in the between-task correlational analysis (see below).

To assess the adaptation shifts, a four-way repeated measures ANOVA was conducted: Condition (/θ/ vs. /s/) × Phase (Baseline ID vs. Adaptation) × Presentation order (/s/ Session 1 and /θ/ Session 3, vs. /θ/ Session 1 and /s/ Session 3) × Experiment (Experiment 1 vs. Experiment 2). A significant main effect was found for Condition (*F* (1, 100) = 568.26, *p* < .001, η2 = .850) as well as for Phase (*F* (1, 100) = 196.55, *p* < .001, η2 = .663). The Condition effect reflects the extremely strong effect of /s/ adaptation. There was no main effect of Presentation order (*F* (1, 100) = .09, *p* = .761, η2 = .001), and no main effect of Experiment, *F* (1, 100) = .02, *p* = .898, η2 < .001). The interaction between Condition and Phase (*F* (1, 100) = 986.33, *p* < .001, η2 = .908) was significant, because the two adaptors shifted the responses in opposite directions. The Condition × Phase × Experiment interaction was not significant, *F* (1, 100) = 1.40, *p* =.240, η2 = .014, showing that the relationship between Condition and Phase was similar in Experiments 1 and 2. Experiment did not interact with any other factors, *p's >.05*.

Natural Accent Accommodation

The natural accent accommodation test included a lexical decision task for the pretest and posttest, with a sentence probe task during training to make sure that the participants listened to the Chinese-accented sentences. Average accuracy on the probe task was 83.7% (*SD*= .15) in Experiment 1 and 86.2% (*SD*=.09) in Experiment 2, indicating that participants did indeed pay attention to the sentences. The pretest/posttest measure was designed to assess whether listening to sentences from a particular Chinese-accented speaker produced any changes in how participants perceived words and non-words from that speaker. Recall that the pilot results provided evidence of adjusting to the accent in a way that increased the degree to which listeners accepted input as words, leading to higher accuracy on words but lower accuracy on non-words. In the pilot tests, this lexical shift was stronger for the critical phonemes (i.e., ones that typically are produced poorly by Chinese speakers) than for other phonemes that are typically not difficult for Chinese speakers.

Following the procedure developed in the pilot testing, average accuracy was calculated for the four types of items for each participant: critical words, critical non-words, filler words, and filler non-words. Then, a difference score for each participant was calculated to index any lexical shift for the critical words – the words containing the critical consonant. This index combined any percentage increase in word report (post critical word - pre critical word) with any percentage decrease in non-word report (post critical non-word – pre critical non-word) by subtracting the latter from the former. This measure thus reflects the overall increase in listeners' tendency to report test items as words. These scores for the critical items were calculated and used in the correlational analysis (to quantify how much listeners adjusted to accents by focusing on accent-dependent sounds); comparable scores were also computed for the filler items.

Figure 4 shows the accuracy results for the pretest and posttest, collapsing across groups that had List A and List B in different orders. As shown in Figure 4, the accuracy of critical words increased from pretest to posttest, whereas the accuracy of critical non-words decreased. A similar pattern was found for filler words and non-words. This pattern was consistent across the two experiments, with two different Chinese speakers.

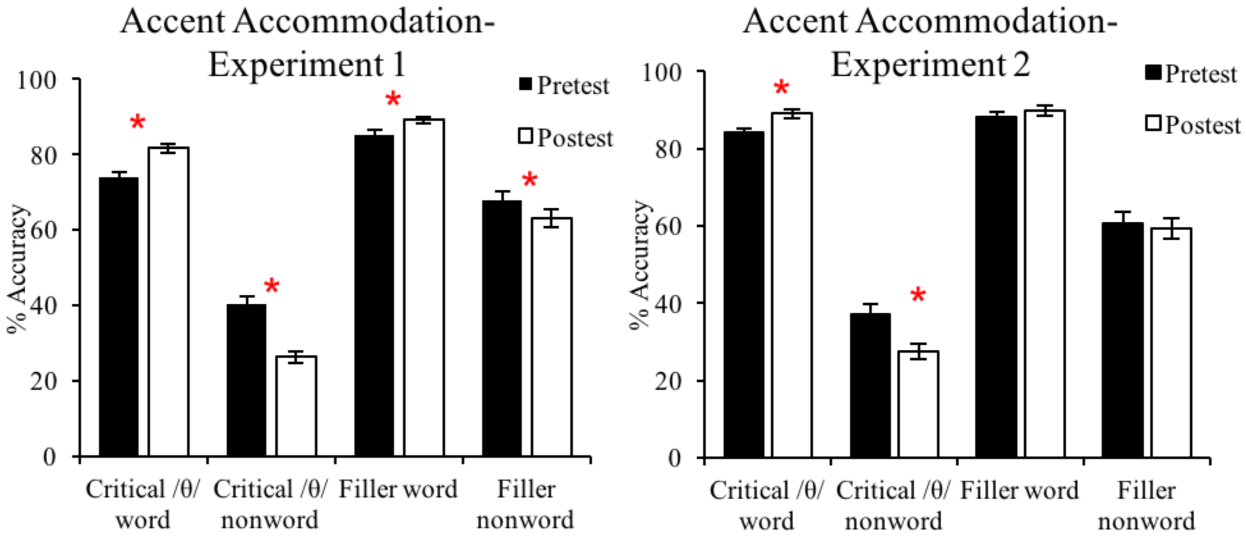


Figure 4. The accuracy of critical and filler words and non-words in the pretest and posttest in Experiment 1 (left panel) and Experiment 2 (right panel). Error bars represent the standard error of the mean.

A five-way repeated measures ANOVA was conducted, with Item (critical vs. filler), Block (pretest vs. posttest), Lexicality (word vs. non-word), Presentation order (pretest List A, posttest List B vs. pretest List B, posttest List A), and Experiment (Experiment 1 vs. Experiment 2) as independent variables and accuracy as the dependent variable. No main effect of Presentation order was found, *F* (1, 100) =.08, *p* =.778, η2 = .001, nor was the main effect of Experiment significant, *F* (1, 100) =1.13, *p* =.291, η2 = .011. The main effect of Item was significant, *F* (1, 100) =627.15, *p* <.001, η2 = .862, reflecting the higher accuracy for filler items than for critical items (which were selected because they included a segment that is difficult for native Chinese speakers to produce accurately). In addition, the main effect of Lexicality was also significant, *F* (1, 100) =513.32, *p* <.001, η2 = .837, with higher accuracy for words than for non-words. The main effect of Block was significant, *F* (1, 100) =3.86, *p* =.052, η2 = .037. The interaction between Block and Lexicality was also significant, *F* (1, 100) =70.76, *p <.001*, η2 = .414, reflecting the fact that word accuracy went up and non-word accuracy went down. The Lexicality × Block × Experiment interaction was significant, *F* (1, 100) =5.04, *p* =.027, η2 = .048, reflecting the somewhat larger effects in Experiment 1 than in Experiment 2.

In Experiment 1, as shown in the left panel of Figure 4, pairwise comparisons (Bonferroni) showed that the accuracy of critical words increased significantly from pretest to posttest (mean difference = 7.9%, *p*<.001) while critical non-word accuracy decreased significantly (mean difference = -13.8%, *p*<.001). The changes for filler words (mean difference = 4.0%, *p*=.005) and for filler non-words (mean difference = -4.8%, *p*=.028) were also significant, though smaller than those for critical items. In Experiment 2 (right panel of Figure 4), pairwise comparisons (Bonferroni) showed that the accuracy of critical words increased significantly from pretest to posttest (mean difference = 4.9%, *p*=.001) while critical non-word accuracy decreased significantly (mean difference = -9.8%, *p*<.001). Note that the larger effects for the non-words are what would be expected if listeners were accommodating to the typical consequences of a Chinese accent. The changes for filler words (mean difference = 1.4%, *p*=.285) and for filler non-words (mean difference = -1.5%, *p*=.584) were not significant.

Correlations of Three Tasks

The primary purpose of the current study is to determine whether individual differences in phonemic category boundary changes are linked to natural accent accommodation – is the mechanism that produces recalibration shifts the same mechanism that is substantially responsible for natural accent accommodation? In order to test if listeners who show big phonemic boundary shifts also show large adjustments on the natural accent accommodation task, a correlational analysis was conducted using the recalibration shifts, the selective adaptation shifts, and the amount of change on the accent accommodation task (which was quantified in the Natural Accent Accommodation section above).

The theoretically most interesting question concerns the relationship between recalibration and accent accommodation: Do listeners improve their understanding of accented speech by shifting their phoneme category boundaries in a way that matches the accent-based phonetic deviation? Recall that this has been a recurring suggestion in the many recent papers that have investigated recalibration. The results of the current study provide no support for this possibility: The correlation between how much participants changed their phonemic boundaries due to recalibration and how much they adjusted to the accent was very weak and non-significant (two-tailed Spearman's correlation in Experiment 1: *r* = -.15, *p*=.306; in Experiment 2: *r* = .17, *p*=.231; two experiments collapsed: *r* = .08, *p*=.426). Figure 5 presents a scatterplot of these data, and makes it clear that there is no systematic relationship.

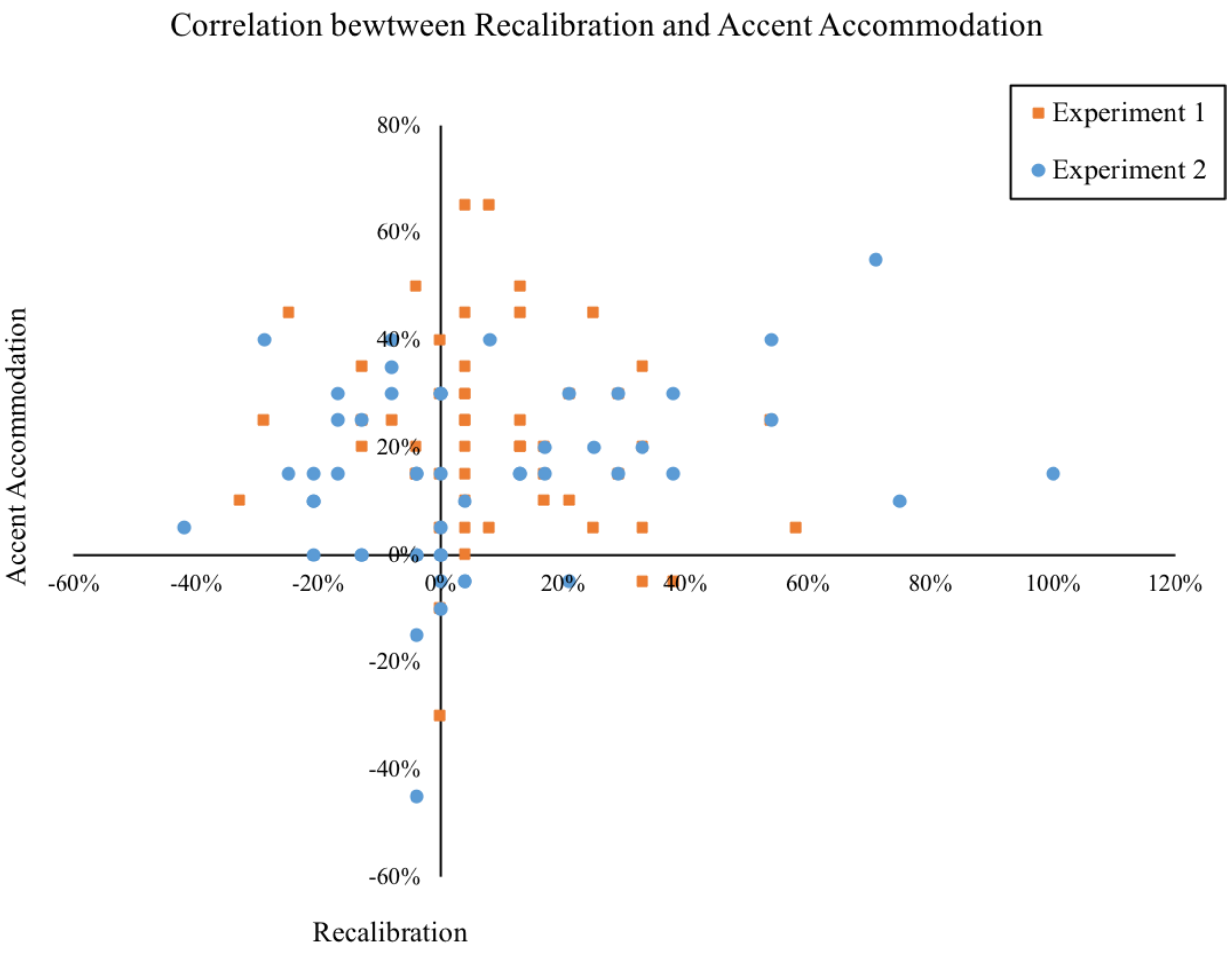


Figure 5. Correlation between recalibration (% /s/ response shift) and natural accent accommodation (% word response shift) in Experiment 1 and Experiment 2.

Given this null effect, the usual concern about a null effect is warranted: Is there really a significant role for boundary shifts in natural accent accommodation that is being missed in the current study? When we described the logic of our study in the Introduction, we laid out three ways that a null effect might be found erroneously: (a) insufficient sample size (power), (b) insufficient between-subject variance, or (c) insufficient stability of one or both effects. Although it is impossible to “prove” a null effect, there are reasons to treat the null effect here as credible.

(a) Insufficient Power: We conducted an *a priori* power analysis and conducted two experiments with sample sizes above those required for a high power level. Thus, the null effect is not plausibly due to insufficient power.

(b) Insufficient Between-Subject Variance: Hedge, Powell, and Sumner (2018) have recently discussed this issue with individual-level measurement. They found low test-retest reliability for several classic tasks, and suggested that it was most likely due to low between-participant variability. If people do not differ much from each other on some measure, then computing individual-level correlations will be problematic. To assess whether this issue might underlie the null correlations, we converted the scores for all of the subjects in Experiments 1 and 2, on each of the three tasks, to z-scores. Figure 6 displays the three distributions. In all three cases, the scores are approximately normally distributed, with substantial between-subject variance. Thus, the null correlations cannot be attributed to there being insufficient between-subject variance.

(c) Insufficient stability of the effects: The results include significant boundary shifts for both recalibration and selective adaptation, with extremely similar results across Experiments 1 and 2. The accent adjustment was significant, was quite similar across Experiments 1 and 2, and closely matched the pattern found in the pilot tests. Thus, at the group level, all three effects were extremely stable. However, for a correlation, such group level stability is not sufficient; individual-level stability is required. For example, if Subject A produces strong recalibration in one test but weak recalibration in another test, while Subject B does the reverse, then correlations of the recalibration scores with another measure will fail – there needs to be a consistent ordering (e.g., Subject A produces small effects consistently, and Subject B produces large effects consistently).

The nature of recalibration (i.e., the inoculation effect that blocks recalibration after exposure to clear tokens) makes it extremely difficult to get a measure of the effect's stability, so it is impossible to completely rule out this possible concern. However, as we noted in the Introduction, if recalibration were to be unstable in this way, it would be quite ill-suited to be the mechanism producing accent accommodation – the intelligibility of accented speech would come and go. We are not aware of any studies assessing this possibility, but it strikes us as unlikely. Moreover, it is not clear how general the concern about between-task correlations really is – robust correlations are still being found between tasks that are arguably more removed from each other than what we tested here. For example, Schmitz et al. (2018) recently reported robust correlations between measures of speech perception and speech production. Perhaps more critically, in a study conducted in the same lab as the current study, with the same participant population, Ishida, Samuel & Arai (2016) found a reliable correlation between two quite different measures of how much different listeners rely on lexical information.

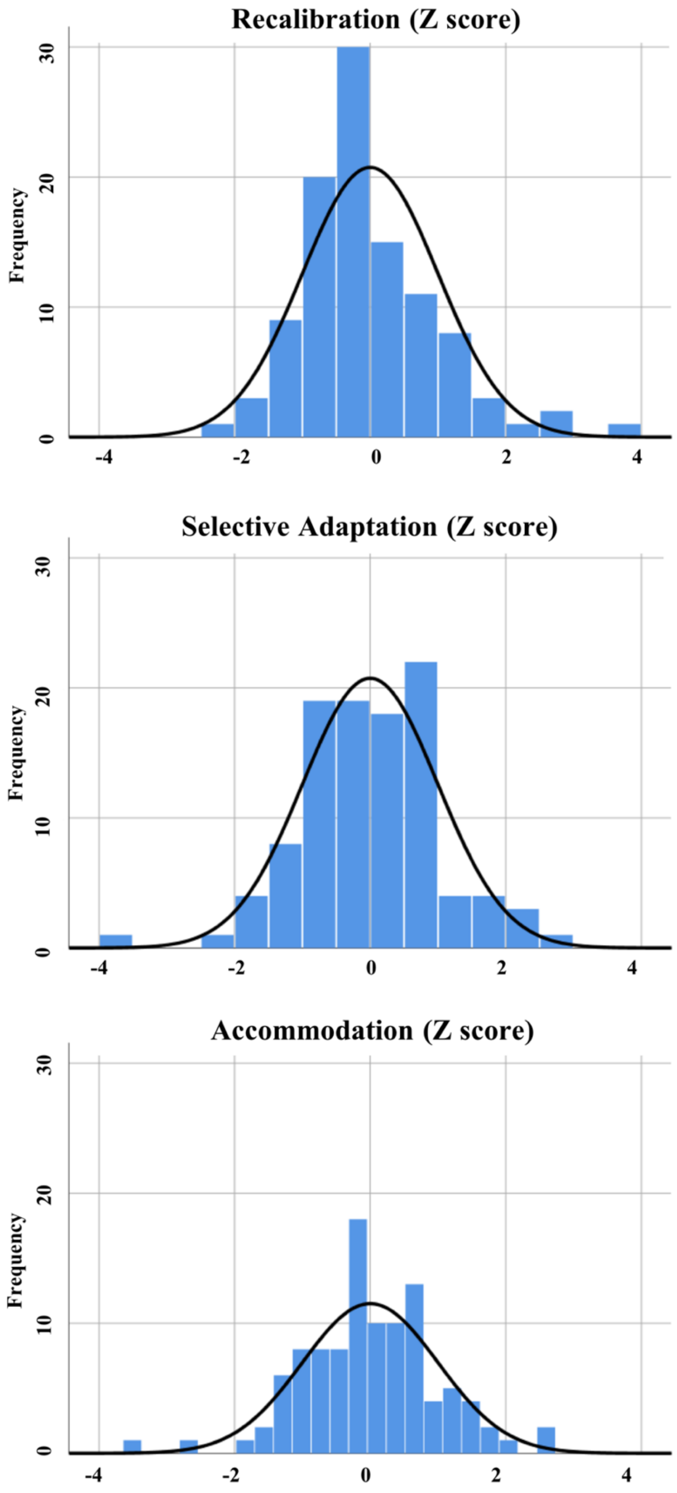


Figure 6. Distribution plots of the three measures (converted to Z scores): recalibration, selective adaptation, and natural accent accommodation.

There was no significant correlation between recalibration and selective adaptation (two-tailed Spearman's correlation in Experiment 1: *r* = .18, *p*=.196; in Experiment 2: *r* = .02, *p*=.903; two experiments collapsed: *r* = .11, *p*=.275). Figure 7 provides the scatterplot for these two measures, and illustrates the lack of any correlation. Similarly, as expected, the shifts caused by selective adaptation were uncorrelated with the degree of accent accommodation (two-tailed Spearman's correlation in Experiment 1: *r* = .02, *p*=.914; in Experiment 2: *r* = .04, *p*=.758; two experiments collapsed: *r* = .07, *p*=.503). Figure 8 provides the scatterplot for these two measures.

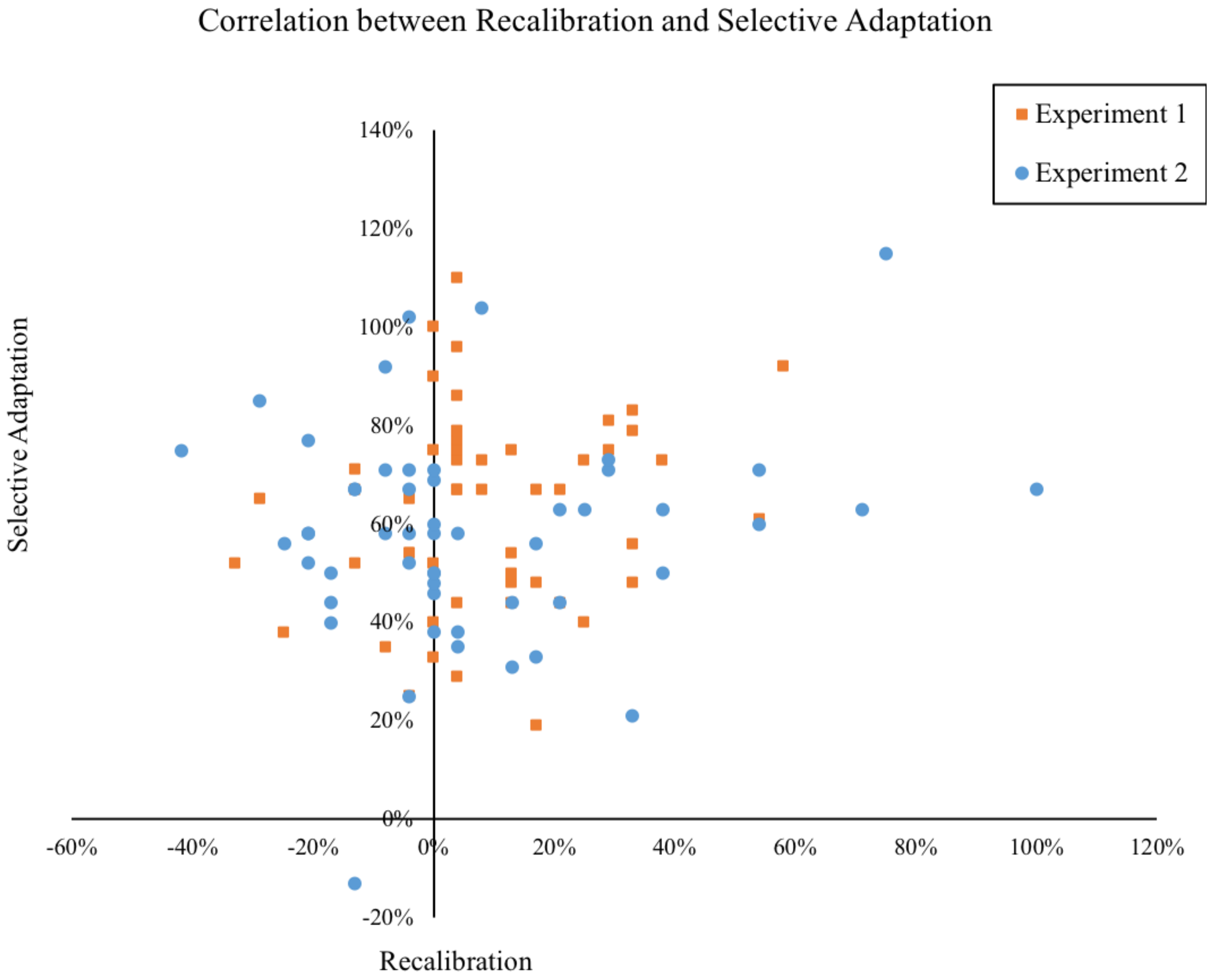


Figure 7. Correlation between recalibration (% /s/ response shift) and selective adaptation (% /s/ response shift) in Experiment 1 and Experiment 2.

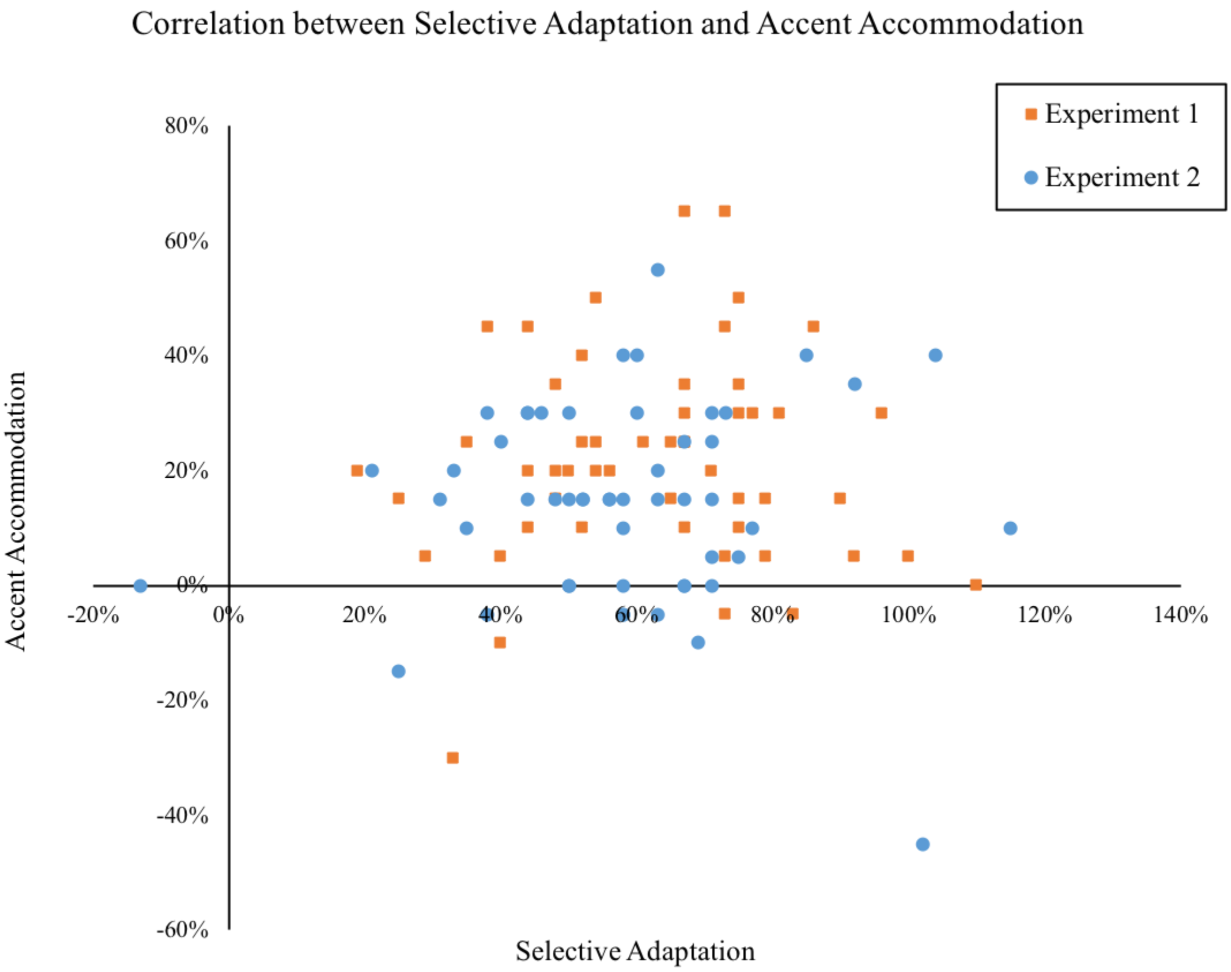


Figure 8. Correlation between selective adaptation (% /s/ response shift) and natural accent accommodation (% word response shift) in Experiment 1 and Experiment 2.

Recall that we saw evidence that recalibration effects in Session 3 were reduced, presumably due to inoculation. To ensure that the null correlations were not a consequence of the inoculation effect on the recalibration task, we conducted an additional correlational analysis that focused on the accented side of the contrast (/θ/), only including participants who did the /θ/ condition in Session 1. The rationale is that if inoculation occurs, the identification functions in Session 3 should be at a baseline level. To be consistent, for selective adaptation, we calculated the shift from baseline for the /θ/ condition, and used these scores for each participant. These new correlational analyses showed the same null results: No significant relationship was found between recalibration and natural accent accommodation, between recalibration and selective adaptation, or between selective adaptation and accent accommodation, all *p's*>.05.

General Discussion

The current study was motivated by the assumption in the recalibration literature that phonemic category boundary change is a major underlying mechanism of natural accent accommodation. We focused on a consonant contrast (/θ/-/s/) that is known to be difficult for Chinese speakers. Experiment 1 and Experiment 2 used the same recalibration and selective adaptation tasks for the /θ/-/s/ contrast, with two different Chinese speakers providing the stimuli used in the accent accommodation task. Across Experiments 1 and 2, the recalibration and selective adaptation effects were very similar, indicating that the effects were reliable and replicable. Participants' performance on the accent accommodation task was also very similar across Experiments 1 and 2.

Boundary Shifts

There is a robust recent literature that shows that listeners retune their phonemic category boundaries as a result of listening to idiosyncratic speech (e.g., Eisner & McQueen, 2005, 2006; Eisner et al., 2013; Kraljic & Samuel, 2005, 2006, 2007; Kraljic et al., 2008; Norris et al., 2003).

Consistent with these findings, in the current study listeners' phonemic category boundaries shifted significantly as a consequence of hearing lexically disambiguated ambiguous phonemes. It is worth noting that we modified the recalibration paradigm in order to measure both sides of each contrast within each individual. As noted before, the recalibration effect is very delicate in certain ways. Anticipating this, the new design was based on extensive pilot testing in an attempt to reduce the inoculation effect of prior exposure to "good" tokens. These procedures seem to have been only partially effective in eliminating inoculation. There was a shift in the usual direction in Session 3, but the recalibration effect was only significant in Session 1, not in Session 3.

The selective adaptation task replicated the well-established finding that after repeated exposure to a stimulus, listeners reduce their report of similar stimuli (Eimas & Corbit, 1973). The effect of selective adaptation was quite robust in both experiments. It is worth noting that the baseline identification was well aligned across the two conditions in each experiment (see Figure 3). This stability demonstrates that preceding each adaptation test with a recalibration test did not interfere with the adaptation measurement; the use of a male voice for recalibration, versus a female voice for selective adaptation, successfully avoided any contamination effects across tasks.

Selective adaptation is usually viewed as being more acoustically driven than recalibration, with the former relying on repeated exposure to a good exemplar of a sound, and the latter on the influence of a lexically-biased context on perception of an ambiguous token. Kleinschmidt and Jaeger (2015) included the two effects in the same computational model, suggesting that they are different manifestations of the same sensitivity to the probability structure of the input. The current study tested these two effects within subjects and found that there was no significant correlation between selective adaptation and recalibration for the /θ/-/s/ consonant contrast. It remains to be seen whether the two phenomena really can be accounted for within a single model.

Accent Accommodation

Various methods have been used to study natural accent accommodation, including priming tasks (Eisner et al., 2013; Weber et al., 2014; Xie & Myers, 2017; Xie, Theodore, & Myers, 2017), transcription tasks (Baese-Berk et al., 2013; Bradlow & Bent, 2008; Gordon-Salant et al., 2010), ERP measures (Romero-Rivas, Martin, & Costa., 2015), and eye-tracking (Dahan et al., 2008). Despite some variation in the time needed for accent adjustment to take place, these studies have consistently shown that natural accent accommodation occurs relatively quickly. The current study was designed to look at adjustment to particular accent markers. Thus, a pretest-training-posttest paradigm was chosen, an approach that has been effective in prior work (e.g., Gass & Varonis, 1984; Wade, Jongman, & Sereno, 2007).

Consistent with previous studies, accent adjustment in the current study occurred relatively quickly, over a 5-min pretest (i.e., lexical decision), a 10-min training period (i.e., the sentence listening task), and a 5-min posttest (i.e., lexical decision, using different test items than the pretest). Listeners showed an increased acceptance of non-words as words on the lexical decision task after training, a pattern that was stable across the two experiments, for both filler items and critical items. Similar results were reported by Maye et al. (2008), who found that accent adjustment involves an increased endorsement of items that were initially not accepted as words due to the accents. In both their study and ours, this increased endorsement rate was used as an index of accent accommodation.

Recall that in the current study, participants provided a rating of their overall familiarity with Chinese accented speech. These ratings can be used to test whether people with more experience with Chinese accented speech show bigger accent accommodation effects. A simple correlational analysis was conducted between familiarity scores and the amount of adjustment for critical items, collapsing across the two experiments. No significant correlation was found between the familiarity ratings and the change in the acceptance of the critical items (two-tailed Spearman's correlation, r = -.12, p =.076). This result is consistent with the talker-specific effects reported in much prior work (e.g., Gass & Varonis, 1984; Jongman, Wade, & Sereno, 2003). It also converges with a recent finding that listeners' experiences with Chinese teaching assistants in classroom settings did not affect their comprehension of speech from a different Chinese speaker in a lab setting (Zheng & Samuel, 2019). Of course, sufficient exposure, to enough accented speakers, can eventually allow listeners to generalize to new accented speakers (e.g., Bradlow & Bent, 2008; Porretta, Tucker, & Jarvikivi, 2016).

Boundary Shifts and Accent Accommodation

The core theoretical question of the current study is whether systematically shifting phoneme boundaries is a primary mechanism in adjusting to accented speech. The recalibration literature has repeatedly shown that listeners shift their phonemic category boundaries after exposure to idiosyncratic speech. Although the ambiguous sounds in these studies are not necessarily identical to accent-based variation, researchers have suggested that recalibration could provide a potential mechanism for natural accent accommodation. Reinisch and Holt (2014) attempted to link the recalibration effect to the natural accent accommodation process by inserting artificially manipulated ambiguous sounds into a naturally accented context. They found that listeners shifted their phonemic boundaries, as in other studies. However, this finding only shows that the recalibration effect was not affected by the global accent; it does not give direct evidence that phonemic category shifting is the basis for accent accommodation. The results from the current study provide no support for such boundary shifts as a substantial basis for accent accommodation, as there was no significant correlation between recalibration and accent accommodation. The absence of evidence is of course not evidence for the absence of an effect, though *a priori* power analyses indicate that the sample size should have been sufficient to find a moderate size effect if one existed, and the stability of the measures across Experiments 1 and 2 was notable. At this point, we believe that any assertion that laboratory-based recalibration effects can explain natural accent accommodation should be supported by evidence that shows such a relationship.

Boundary Shifts vs. Criteria Relaxation

Within the natural accent accommodation task itself, the participants consistently showed an increased endorsement rate (as words) for both the critical items and for the filler items. The critical items were selected to reflect a particular contrast that Chinese speakers have difficulty with: Chinese speakers tend to produce /θ/ as more /s/-like. Learning these particular accent markers led to increased "word" responses for both critical words (i.e., increased accuracy) and critical non-words (i.e., decreased accuracy). Interestingly, because the filler items were selected to avoid sounds that are difficult for Chinese native speakers, fillers items did not strongly reflect pronunciations characteristic of Chinese accents, yet the same pattern (although smaller than for the critical items) was observed. The results suggest that listeners were consistently more tolerant of non-standard pronunciations as a result of the accent accommodation process.

These results raise the question of whether natural accent accommodation involves some type of criteria relaxation. As noted earlier, criteria relaxation involves a tolerance for irregular speech patterns from accented speakers, with more flexibility about what is acceptable. The results from the current study suggest that listeners tend to be more accepting of all sounds from the accented speaker, whether oddly-produced or not. This pattern is reminiscent of the findings of Hanulíková and colleagues (2012), in the syntactic processing of non-native speech. They found that the P600 (a neural correlate of syntactic violations) was generated when listeners heard a native speaker produce a syntactic violation, while the same violation by a non-native speaker did not yield a P600. This result indicates that listeners were more willing to accept syntactic errors from a non-native speaker than from a native speaker, a relaxation of criteria for non-native speech.

Reinisch and Weber (2012) also found that listeners adjusted to lexical stress errors in accented speech, indicating that natural accent adjustment occurs not only at the syntactic level but also at the suprasegmental level. Witteman, Weber, and McQueen, (2013) used a cross-modal priming task to investigate listeners' adjustments to accented words that contained small or large vowel deviations from standard pronunciations, and found that even without specific training, adults accommodated to test items that contain large pronunciation deviations. In addition, White and Aslin (2011) tested toddlers' accommodation to accents in which specific vowels were changed (e.g., "dog" produced as /dæg/), and found that listeners not only accommodated to new items with the same mispronunciations (/sæk/ for "sock") but also to phonemes with a different mispronunciation that was acoustically similar to the standard pronunciation (/sek/ for "sock"). Collectively, these studies provide evidence to support the criteria relaxation idea. It is plausible that when confronted with irregular patterns in accented speech, listeners broaden what they will accept as a word.

As we noted in the Introduction, although there are many findings in the literature that suggest an important role for criteria relaxation in adjusting to accented speech, there are a number of results that are not well accounted for through this mechanism. Recall that Maye et al. (2008) created an accent by lowering front vowels and keeping everything else normal. Accent adjustment only occurred in one direction, rather than the bidirectional change implied by a general relaxation. Bradlow and Bent (2008) did not find any improvement in listeners' understanding of Slovakian-accented English after exposure to a set of Chinese-accented English speakers, showing no evidence of more tolerance for all deviations, even from a different accent. In Xie and Myers's (2017) study, listeners were exposed to one or several Chinese-accented English speakers, giving them experience with many non-native tokens of final /d/. On a test following this exposure, Xie and Myers found no improvements in listeners' perception of /t/-final words, as might have been expected if the listeners had undergone a general relaxation of the criterion for deciding whether a sound was /d/ or /t/. Clearly, a criterion relaxation mechanism cannot account for all accent accommodation – multiple mechanisms are presumably involved.

One possible additional mechanism is potentially available with certain types of accented speech. Weber et al. (2014) suggested that if a certain sound is consistently produced as a different sound (i.e., the "bad map" case that Sumner, 2011, discussed) listeners could learn to map both sounds to appropriate lexical items. They studied Italian-accented English in which /I/ is typically produced as /i/ (for example, "Italy" would be said as "Eataly"). Listeners could map both /I/ and /i/ to lexical representations that would typically have /I/, allowing both "Italy" and "Eataly" to activate the lexical representation for "Italy". Samuel and Larraza (2015) examined a comparable case for Spanish-accented Basque productions of affricates, and found evidence for what they called a “dual mapping” of segments to the lexicon. This dual-mapping account may explain some of the accommodation that occurs to accented speech, especially in cases in which the accent reliably produces a segment that is also available in the listener’s native language. When the accent does not produce such an alternative segment (either because the change is only partial, or because the change does not map to an existing L1 alternative), the criterion relaxation or boundary shift accounts may be more appropriate.

We noted above that the null correlation we found between accent accommodation and recalibration was not due to a lack of power, or to a lack of between-subject variability. Each of the two experiments included a sufficient sample to find a moderate-size effect if there was one to be found. The broad and normal distribution of the scores on each task, shown in Figure 6, was complemented by group-level results that were extremely stable. This stability was evident in the very similar patterns shown in the two panels of each data figure: Across Experiments 1 and 2, the results were strikingly consistent for all three tasks (i.e., recalibration, selective adaptation, and accent accommodation).

We noted that the one potential concern about the correlations that we could not statistically reject was that at the individual-subject level, recalibration might not be stable enough to produce reliable correlations. The recent concern about correlations of individual-difference measurement (Hedge et al., 2018; Paap & Greenberg, 2013) does call for a cautious response to the patterns we observed. A recent study by Saltzman and Myers (2018) suggests that this concern must be taken seriously. Saltzman and Myers conducted a recalibration experiment in which they measured each participant’s recalibration effect twice, across sessions held 5-11 days apart. In their study, at the group level, the size of the recalibration effect was almost identical across the two sessions. However, when they correlated the size of the shifts on a per-subject basis, the reliability was very low (*r* =.12), and not significant.

The concerns about the stability of individual-level measurement are beginning to emerge in a number of domains. For example, in the last few years, people working on the “bilingual advantage” in executive control have been finding very low correlations between tasks that supposedly tap very similar attentional mechanisms (Fan, Flombaum, McCandliss, Thomas, & Posner, 2003; Humphrey & Valian, 2012; Keye, Wilhelm, Oberauer, & Van Ravenzwaaij, 2009; Kousaie & Phillips, 2012; Stins, Polderman, Boomsma, & de Geus, 2005; Paap, & Greenberg, 2013). The flanker task and the Simon task are both used to index inhibitory control, but Paap and Greenberg (2013) found a very low correlation between them (*r* = -.01).

Given these general concerns about between-task correlations, the current results should be considered with appropriate caution. At the same time, as we noted above, there are both empirical and logical reasons to take our results seriously. Empirically, correlational tests are still succeeding in many studies (e.g., Schmitz et al., 2018), notably including a recent study from the same lab as the current study, with the same participant population (Ishida, Samuel, & Arai, 2016). Logically, as we noted in the Introduction, if recalibration is too unstable to support reliable correlations, it would be too unstable to support natural accent accommodation.

Conclusion

The two experiments in the current study constitute the first attempt to look at the relationship between low-level boundary adjustment effects (recalibration and selective adaptation) and higher order accommodation of natural accents. The results provide no support for the view that recalibration of phonemic boundaries plays a central role in natural accent accommodation. Instead, the outcome for the accent accommodation task is more consistent with the view that at least part of accent accommodation involves a relaxation of phonemic categorization criteria in the word recognition process.

Even if our negative results are interpreted cautiously, they highlight the need for some affirmative evidence to support the suggestion (e.g., Eisner, Melinger, & Weber 2013; Kraljic & Samuel, 2006; Mitterer & McQueen, 2009) that the boundary shifts found in studies of recalibration can provide a mechanism for accommodating to accented speech. We have provided an initial test of this often-made claim, and have obtained a negative result. Only when multiple additional tests have been conducted will we know whether there really is a strong connection between recalibration and adjusting to accented speech. We hope that researchers will accept this challenge, and provide these critical additional tests.

Acknowledgements

Support was provided by Ministerio de Ciencia E Innovacion, Grant PSI2017-82563-P, Centro de Excelencia Severo Ochoa, Grant SEV-2015-0490, by the Basque Government through the BERC 2018-2021 program, and by the National Science Foundation under Grant IBSS-1519908. We thank Antonio Freitas, Richard Gerrig, and Marie Huffman for their constructive suggestions. We also thank James M. McQueen for valuable comments on a previous version of the paper.

References

Bamford, J., & Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. Speech-hearing tests and the spoken language of hearing-impaired children, 148-187.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk after effect. *Psychological Science*, *14,* 592-597.

Bradlow, A. R. & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106* (2), 707-729.

Brannen, K. (2002). The role of perception in differential substitution. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *47(1-2),* 1-46.

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.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116(6),* 3668-3678.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*(3), 710-718.

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

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, 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.

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

Fan, J., Flombaum, J. I., McCandliss, B. D., Thomas, K. M., & Posner, M. I. (2003). Cognitive and brain consequences of conflict. *Neuroimage*, *18(1)*, 42-57.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25(4)*, 437-470.

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.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65-87.

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

Hancin-Bhatt, B. J. (1994). Phonological transfer in second language perception and production (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Hanulíková, A., & Weber, A. (2012). Sink positive: Linguistic experience with th substitutions influences nonnative word recognition. *Attention, Perception, & Psychophysics*, *74(3)*, 613-629.

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24(4)*, 878-887.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166-1186.

Ishida, M., Samuel, A.G., & Arai, T. (2016). Some people are "more lexical" than others. *Cognition*, *151*, 68-75.

Humphrey, A. D., & Valian, V. V. (2012). Multilingualism and cognitive control: Simon and flanker task performance in monolingual and multilingual young adults. In 53rd Annual meeting of the Psychonomic Society.

Jongman, A., Wade, T., & Sereno, J. (2003). On improving the perception of foreign-accented speech. In Proceedings of the 15th international congress of phonetic sciences (pp. 1561-1564).

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252-1263.

Kent, R. D., Read, C., & Kent, R. D. (1992). *The acoustic analysis of speech* (Vol. 58). San Diego: Singular Publishing Group.

Keye, D., Wilhelm, O., Oberauer, K., & Van Ravenzwaaij, D. (2009). Individual differences in conflict-monitoring: testing means and covariance hypothesis about the Simon and the Eriksen Flanker task. *Psychological Research PRPF*, *73(6)*, 762-776.

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.

Kousaie, S., & Phillips, N. A. (2012). Conflict monitoring and resolution: Are two languages better than one? Evidence from reaction time and event-related brain potentials. *Brain Research*, *1446*, 71-90.

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

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.

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one*, *4*(11), e7785.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66(2)*, 232-258.

Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, *58*, 1-21.

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.

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

Rogers, C. L., & Dalby, J. (2005). Forced-choice analysis of segmental production by Chinese-accented English speakers. *Journal of Speech, Language, and Hearing Research*, *48*(2), 306-322.

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, *9*, 167.

Saltzman, D., & Myers, E. (2018). Listeners are maximally flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, *25*(2), 718-724.

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

Samuel, A.G., & Larraza, S. (2015). Does listening to non-native speech impair native speech perception? *Journal of Memory and Language*, *81*, 51-71.

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science*, *15(6)*, 732-738.

Schmitz, J., Diaz, B., Fernandez Rubio, K., & Sebastian-Galles, N. (2018). Exploring the relationship between speech perception and production across phonological processes, language familiarity, and sensory modalities. *Language, Cognition and Neuroscience*, *33*, 527-546.

Stins, J. F., Polderman, J. C., Boomsma, D. I., & de Geus, E. J. (2005). Response interference and working memory in 12-year-old children. *Child Neuropsychology*, *11(2)*, 191-201.

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

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.

Weber, A., Di Betta, A. M., & McQueen, J. M. (2014). Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners. *Journal of Phonetics*, *46*, 34-51.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372-384.

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.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalization of accept adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*, 196-210.

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., 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.

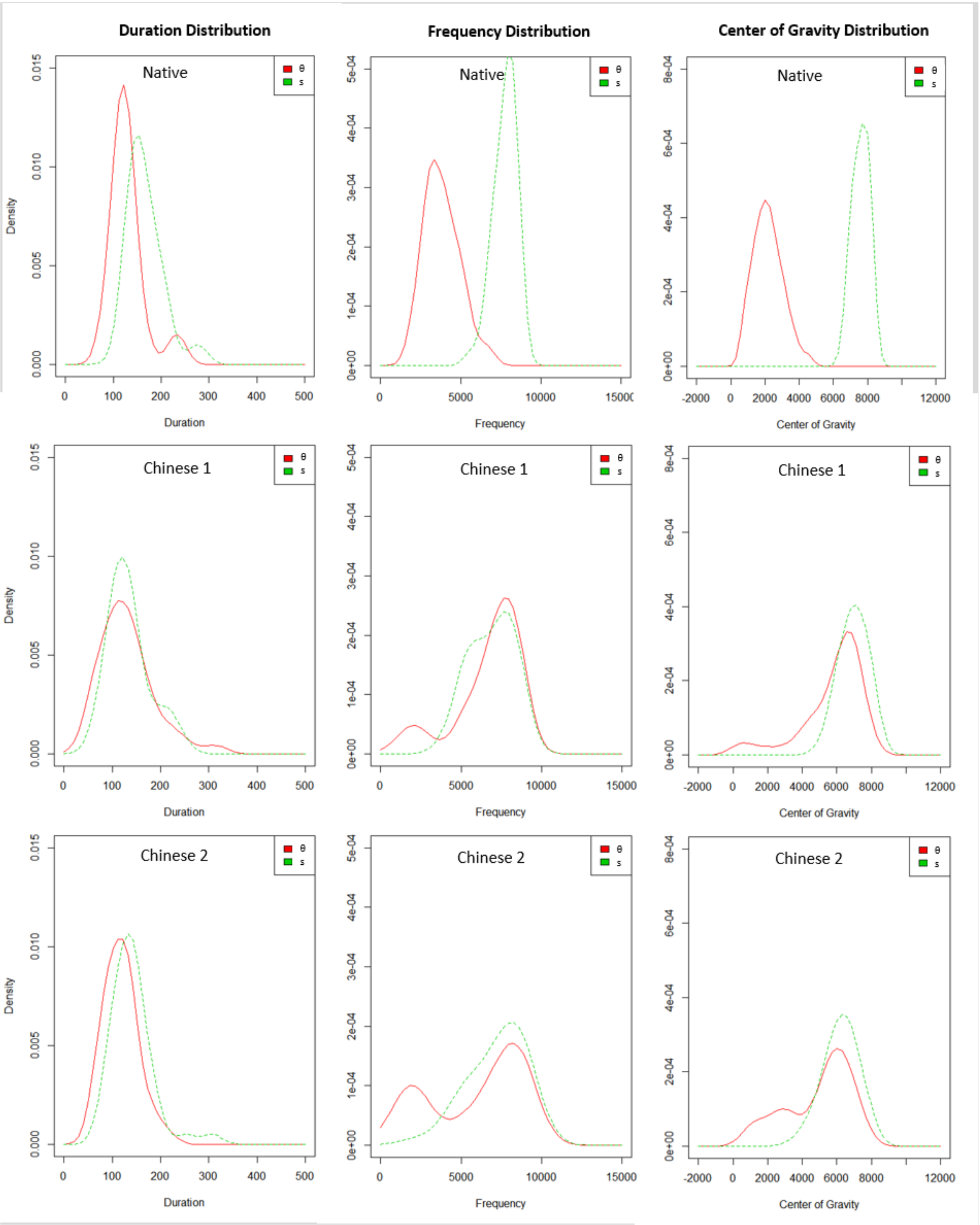
Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 200.

Zheng, Y., & Samuel, A. G. (2019). How much do visual cues help listeners in perceiving accented speech? *Applied Psycholinguistics*, *40(1)*, 93-109.

Zheng, Y., & Samuel, A. G. (2017). Does seeing an Asian face make speech sound more accented? *Attention, Perception, & Psychophysics*, *79*(6), 1841-1859.

Appendix 1

Density plots of native pronunciation and two Chinese speakers for the /θ/-/s/ contrast



Appendix 2

Critical Words, Fillers, and Non-words for /θ/-/s/ Contrast on the Recalibration Task

Critical /θ/ words	**Ambiguous step**		**Non-words**	
urethane	30%/s/70%/θ/	tomorrow	molorat	omblatalal
amphitheater	35%/s/65%/θ/	opener	leggarer	borrorana
anything	50%/s/50%/θ/	polymer	anguilder	demureanal
apathetic	25%/s/75%/θ/	auditor	connontnoor	brahartacko
apathy	30%/s/70%/θ/	commoner	becoor	dudanaco
dorothy	35%/s/65%/θ/	guttural	dentakter	andegaul
empathy	25%/s/75%/θ/	perusal	lampunger	bocowlable
everything	30%/s/70%/θ/	onlooker	elembre	polacual
hypothetical	20%/s/80%/θ/	pagoda	anoda	bonmalatad
jonathan	40%/s/60%/θ/	gunpowder	oggander	ampoterate
marathon	35%/s/65%/θ/	determine	combeter	odecogo
neanderthal	40%/s/60%/θ/	untoward	adgendoy	horabtalane
polyethylene	25%/s/75%/θ/	awarded	entonker	ancorrackant
telepathy	25%/s/75%/θ/	moderate	guncore	bonconkartat
timothy	35%/s/65%/θ/	kimono	cumpamer	memebable
unauthentic	35%/s/65%/θ/	durable	pleophe	altartalized
Critical /s/ words		optional	akelen	rapombargad
admissible	25%/s/75%/θ/	opportune	cleniot	radorcattoon
ambassador	20%/s/80%/θ/	murderer	bulerame	oudrenoa
coliseum	15%/s/85%/θ/	abortion	booktugner	morachable
contraceptive	10%/s/90%/θ/	monitored	otler	omblegontac
democracy	15%/s/85%/θ/	maternal	haderate	reifonairo
dinosaur	10%/s/90%/θ/	undertow	pocorome	ancarruntlo
episode	20%/s/80%/θ/	topical	annantor	omparkandar
eraser	20%/s/80%/θ/	modulation	etoced	altulable
hallucinate	10%/s/90%/θ/	carbohydrate	coerpage	caltacater
inconclusive	15%/s/85%/θ/	hydrocarbon	etugant	premetetor
legacy	15%/s/85%/θ/	coordinate	cayarac	dadargora
medicine	15%/s/85%/θ/	adorable	negryhad	motounalad
parasite	15%/s/85%/θ/	computation	pobtler	dadadratar
participate	15%/s/85%/θ/	marketable	collattar	contaluow
pregnancy	10%/s/90%/θ/	mutilated	nannotad	dandarallad
reconcile	25%/s/75%/θ/	dominated	kedoac	amahaate
Fillers		termination	pelayde	odanatar
domination		compilation	anapte	andarkackood
multitude		undertaker	altercole	mypuroucly
terminated		colorado	nererant	dynremacal
monopolize		aluminum	dioryle	hapabutda
mutilation		underwater	meloded	reatonape
workable		congregation	relecker	prodabanga
coronation		commendation		

Appendix 3

Words and Non-words for /θ/-/s/ Contrast on the Accent Accommodation Task

Critical word-list A	Critical non-word-listA	Filler word -listA	Filler non-word -listA
withdraw	wishold	trouble	booknark
tooth	wealsy	calendar	subben
thief	ausentic	eyebrow	uncerpain
underneath	toospick	program	gutterfly
health	healsy	punch	najor
pathetic	soughtfully	awake	headabe
third	serapy	sweater	cirdle
enthusiasm	caserine	round	biagram
nothing	seorist	robot	bapyard
athletic	ausority	flexible	brocessor
birthday	ensusiastic	screw	nanual
pathology	sanksgiving	grateful	teanut
synthetic	esical	crowd	callot
sympathy	orsography	colorful	sbratch
theft	aslete	tolerate	skeak
unfaithful	mesod	sponsor	lezel
therapist	sirteen	battle	gatch
authorize	freesinker	download	flagile
forthcoming	seater	notebook	bolor
thankful	lengsening	joke	ganana
withhold	wisdraw	bookmark	troudle
wealthy	toos	sudden	palendar
authentic	sief	uncertain	eyekrow
toothpick	underneas	butterfly	probram
healthy	heals	major	dunch
thoughtfully	pasetic	headache	awape
therapy	sird	circle	sweaper
catherine	ensusiasm	diagram	rounk
theorist	nosing	backyard	rokot
authority	asletic	processor	flexitle
enthusiastic	birsday	manual	sbrew
thanksgiving	pasology	peanut	drateful
ethical	synsetic	carrot	crowk
orthography	sympasy	scratch	tolorful
athlete	seft	steak	polerate
method	unfaisful	level	skonsor
thirteen	serapist	catch	bapple
freethinker	ausorize	fragile	townload
theater	forscoming	color	nokebook
lengthening	sankful	banana	jode

Appendix 4

Sentences on the Accent Accommodation Task (Words in Italics are Related to the Probe)

Practice

sentence1	sentence2	sentence3	probe	Ans
The boy hurried to school.	She argued with her sister.	A girl came into the room.	The boy rushed to class.	yes
She is waiting for her bus.	*The girl caught a cold.*	The police chased the car.	The girl was sick.	yes
The girl played with the baby.	Lemons grow on trees.	*She is washing her dress.*	She is cleaning her shoes.	no

Formal

sentence1	sentence2	sentence3	probe	Ans
They had a lovely day.	The old gloves are dirty.	The raincoat is very wet.	They had a great day.	yes
He needed his vacation.	A girl kicked the table.	School finished early today.	He needed some time off from work.	yes
The machine was very noisy.	The driver waited by the corner.	The wife helped her husband.	The machine was making a lot of noise.	yes
The truck carried fruit.	The janitor swept the floor.	He found his brother.	The truck carried healthy food.	yes
The two children are laughing.	A friend came for lunch.	The matches are on the shelf.	The two kids are having fun	yes
The teapot is very hot.	The shoes were very dirty.	The five men are working.	The kettle is cold.	no
The truck drove up the road.	The ground was very hard.	The cow gave some milk.	The truck broke down.	no
Father looked at the book.	The rain came down.	She drinks from her cup.	Father is burning his books	no
The match fell on the floor.	The young boy left home.	A sharp knife is dangerous.	The lighter fell on the floor.	no
The plant is hanging above the door.	*The faucets are above the sink.*	The house had a nice garden.	The hot and cold handles are over the sink.	yes
The young people are dancing.	*The driver started the engine.*	The match boxes are empty.	The driver turned on the motor.	yes
The driver lost his way.	*Snow falls at Christmas.*	They had two empty bottles.	There is snow at Christmas.	yes
The jug is on the shelf.	*The car engine is running.*	They say some silly things.	The auto's motor is on.	yes
The boy is running away.	*They are buying some bread.*	A cat jumped off the fence.	They are getting some food	yes
The orange was very sweet.	*The football game is over.*	The fire was very hot.	The orange was rotten.	no
Sugar is very sweet.	*They went on a vacation.*	The scissors are very sharp.	They went to work.	no
Some men shave in	*The train is moving*	The milkman	The bus traveled in	no

the morning.	*fast.*	drives a small truck.	traffic.	
The police helped the driver.	*The oven door was open.*	Men wear long pants.	The oven was burning hot.	no
The dog played with a stick.	*The family likes fish.*	The farmer keeps a bull.	The family hates seafood	no
The cat drank from a saucer.	*He cut his finger.*	The floor looked clean.	He cut watermelons	no
Father paid at the gate.	*The cat is sitting on the bed.*	The mailman brought a letter.	The cat was frightened by the storm.	no
The clown had a funny face.	The ball is bouncing very high.	*The girl has a picture book.*	The girl owns a photo book.	yes
She spoke to her son.	The train had a bad accident.	*The child grabbed the toy.*	The kid got the toy.	yes
Some sticks were under the tree.	She had her spending money.	*The mailman shut the gate.*	The postal carrier closed the gate	yes
A fish swam in the pond.	He is washing his face.	*The little baby is sleeping.*	The little infant is resting.	yes
The car is going too fast.	The dog drank from a bowl.	*The children are walking home.*	The kids are going home.	yes
The family bought a house.	Bananas are yellow fruit.	*Milk comes in a carton.*	Milk arrives in a container.	yes
The children waved at the train.	The bus left early.	*The sky was very blue.*	The sky looked stormy.	no
The oven is too hot.	The dog came back.	*The dishcloth is very wet.*	The dishcloth fell on the ground.	no
The dog made an angry noise.	The glass bowl broke.	*The house had nine rooms.*	The house was fairly small	no
She cut with her knife.	The jelly jar was full.	*They ate the lemon pie.*	They tasted the cherry pie.	no
The ball went into the goal.	The father is coming home.	*The green tomatoes are small.*	The tomatoes are ripe	no

Appendix 5

Language Experience Questionnaire

Name Code: _________

(1) Date: ____________

(2) Age: _____________

(3) Gender: __________

(4) Place of birth (city, country): ____

(5) Native language(s): ____

(6) Please list all the languages you know in order of dominance. Then rate your proficiency in each of the languages. 0=cannot speak. 100= native speaker.

0 10 20 30 40 50 60 70 80 90 100

Native language: _________________

Language 2: ______________________

Language 3: ______________________

(7) What other countries you have lived/studied in, and for how long?

____

(8) Have you studied any languages (other than your native language) in school? If so, for how long? ____

(9) We are interested in your life experience with foreign accents. On a scale from 0-10, please rate how much you have experienced Chinese-accented English.

Not familiar at all Very familiar

0 1 2 3 4 5 6 7 8 9 10