Bad maps may not always get you lost: Lexically-driven perceptual recalibration for substituted phonemes.

Jeanne Charoy

Department of Psychology, Stony Brook University

Arthur G. Samuel

Department of Psychology, Stony Brook University

Basque Center on Cognition Brain and Language, Donostia-San Sebastian 20009 Spain

IKERBASQUE, Basque Foundation for Science, Bilbao 48011 Spain

Correspondance concerning this article should be addressed to Jeanne Charoy

Contact: jeanne.charoy.collilieux@gmail.com

Results were presented in November 2021 at the 61st Annual Meeting of the Psychonomic Society (virtual).

Data and analyses can be found at https://github.com/jeanne-charoy/BadmapProject

Abstract

The speech perception system adjusts its phoneme categories based on the current speech input and lexical context. This is known as lexically-driven perceptual recalibration, and it is often assumed to underlie accommodation to non-native accented speech. However, recalibration studies have focused on maximally ambiguous sounds (e.g., a sound ambiguous between "sh" and "s" in a word like "superpower"), a scenario that does not represent the full range of variation present in accented speech. Indeed, non-native speakers sometimes completely substitute a phoneme for another, rather than produce an ambiguous segment (e.g., saying "shuperpower"). This has been called a "bad map" in the literature. In this study, we scale up the lexically-driven recalibration paradigm to such cases. Because previous research suggests that the position of the critically accented phoneme modulates the success of recalibration, we include such a manipulation in our study. And to ensure that participants treat all critical items as words (an important point for successful recalibration), we use a new exposure task that incentivizes them to do so. Our findings suggest that while recalibration is most robust after exposure to ambiguous sounds, it also occurs after exposure to bad maps. But interestingly, positional effects may be reversed: recalibration was more likely for ambiguous sounds late in words, but more likely for bad maps occurring early in words. Finally, a comparison of an online versus in-lab version of these conditions shows that experimental setting may have a non-trivial effect on the results of recalibration studies.

The speech signal is notoriously variable: The same word is produced with different acoustic properties depending on factors such as the speaker (e.g., different pitches, accents, speech rates…), environment (e.g., a church, a busy street…), linguistic context (e.g., coarticulation), or speech style (e.g., casual, formal). Even when an acoustic cue can reliably distinguish two sounds for one speaker, it may not do so between speakers, creating ambiguity (e.g., Newman, Clouse, & Burnham, 2001). This is the well-known problem of *lack of invariance* in speech: how are variable acoustic signals mapped onto the same abstract linguistic percept?

The system responds to this problem, in part, by maintaining enough flexibility to accommodate to inputs that deviate from long-term representations. One well-known phenomenon that may stem from this need for flexibility is lexically-driven perceptual recalibration, the finding that boundaries between phoneme categories can be shifted given exposure to lexical stimuli that include ambiguous segments (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007; Norris, McQueen, & Cutler, 2003). The phenomenon was first demonstrated by Norris et al. (2003), who presented listeners with words that contained a phoneme designed to be perfectly ambiguous between /s/ and /f/ (where ambiguity is defined by pretests with native listeners). Importantly, the words were chosen so that only one interpretation of the ambiguous phoneme was possible. For example, one group would only hear it in lexical contexts that required /s/. Later, when those listeners were asked to categorize items on an /s/-/f/ continuum, they were more likely to categorize ambiguous phonemes as /s/. A different group of listeners heard the ambiguous segments in lexical contexts that required /f/, and these listeners subsequently identified more members of the continuum as /f/. The boundary between the

listeners' /s/ and /f/ categories had shifted to reflect what they had learned about the pronunciation of these phonemes in the speech they had been exposed to.

Lexically-driven perceptual recalibration is often assumed to be involved with another well-known phenomenon: Listeners' ability to accommodate to non-native accented speech, as shown by improvements in comprehension (Baese-Berk, Bradlow, & Wright, 2013; Bent & Bradlow, 2003; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Porretta, Tucker & Jarvikivi, 2016; Witteman, Weber, & McQueen, 2010, 2013, 2014; Witteman et al., 2015). For example, Bradlow and Bent (2008) note that "talker dependent adaptation to foreign accented speech […] is consistent with the general idea of lexically-driven perceptual learning for speech as described in Norris et al. (2003) and subsequent studies" (p.13). Indeed, it is easy to link the two intuitively. In accented speech, phonemes are produced in ways that depart significantly from prototypical forms due to transfers between the speaker's first and second languages. If the system is able to readjust its phoneme categories to match these departures, we would expect increases in comprehension (as observed in the accent accommodation literature). However, the relationship between lexically driven perceptual recalibration and accommodation to accented speech remains to be established. In fact, the currently available evidence does not support a link between the two (Babel, Johnson, & Sen, 2021; Zheng & Samuel, 2020).

A limitation in assessing the possible relationship is that the perceptual recalibration literature has mostly focused on a simplified scenario: One perfectly ambiguous phoneme embedded in an otherwise clear, native sounding disambiguating word, typically located in the middle or end of that word (following Norris et al., 2003's original design). While this

simplification allows for control over the accentedness of the phonemes and lets us observe phoneme-specific recalibration, it does not reflect the range of experiences a listener gets with natural speech, especially accented speech. There are at least three important ways in which natural accented speech may deviate from this scenario. First, accented phonemes may occur anywhere in a word, not just in medial or final position. A handful of studies have investigated how this may affect recalibration (Jesse & McQueen, 2011; Samuel, 2016). For example, Jesse and McQueen (2011) used a classic recalibration paradigm but manipulated the position of the critical ambiguous phoneme at exposure and test, such that it could either occur at onset or coda (resulting in four conditions: onset-onset, onset-coda, coda-coda, coda-onset). They found that when distorted sounds occur early in a word, before lexical access can disambiguate them, recalibration effects tend to be small or absent. This suggests that such effects may rely on a specific set of conditions to be optimal.

Second, accented phonemes usually do not occur within otherwise native sounding speech, but rather within a global accent affecting several segments. This was addressed in one study that found that a globally accented context does not block recalibration for artificially accented ambiguous phonemes (Reinisch & Holt, 2014). Therefore, recalibration would not be limited by this particular aspect of natural accented speech, although further investigation is needed to establish whether this holds for different accents and, importantly, accent strength (in their study, Reinisch & Holt used weak to medium accented English). Importantly, these results only show that recalibration for one ambiguous phoneme is not blocked in more broadly accented speech; they do not speak to what role recalibration may play, if any, in adjusting to accented speech.

This leads to the third way in which perceptual recalibration studies may oversimplify reality: accented phonemes in natural speech may not be perfectly ambiguous, but rather fall anywhere on the continuum from prototypical to fully mispronounced. The current paper focuses on this latter scenario: Is recalibration observed after exposure to phonemes that clearly fall into the wrong category? To position our experiments, we first aim to get a clearer picture of the nature of mispronunciations in accented speech by reviewing studies that evaluated corpora of recorded accented speech.

We are not aware of any systematic reviews of the distribution of segmental variation in accented speech, i.e., how often do non-native speakers produce phonemes in an atypical way, and how atypical are those productions when they occur (i.e., is it fair to assume they are mostly ambiguous ?). However, many individual studies have reported how the production of particular phonemes differs between native speakers and non-native speakers of a language or dialect (Best et al., 2015; Bion, Escudero, & Morrison, 2008; Burgos, et al., 2014; Cebrian, 2007; Cutler, Smits, & Cooper, 2005; Dufour, 2007; Evanini & Huan, 2012; Fabra & Romero, 2012; Flege, 1987; Bohn & Flege, 1992; Flege, Munro, & Skelton, 1992; Hanulíková & Weber, 2012; Jia, et al., 2006; Levy & Law, 2010; Lombardi, 2003; Rau, Chang & Tarone, 2009; Rogers & Dalby, 2005; Tsukada,et al., 2005; Zhang & Xiao, 2014). Phoneme productions are usually evaluated in one of three ways (sometimes in combination): 1) trained phoneticians transcribe the recorded phonemes; 2) groups of naïve native speakers judge the phonemes, indicating what sounds they hear; or 3) acoustic analyses are conducted on the recordings. We limited our review to studies that relied on listeners' judgments (either phoneticians or naïve listeners) because our focus is on

listeners' perception of accented speech; patterns observed in acoustic analyses, while clearly important, are not always consistent with perceptual results (see Bohn & Flege, 1992).

Phonemes that are likely to cause difficulties for non-native speakers are those that do not exist in their first language. A classic example is the English /l/-/r/ contrast that distinguishes words like "lock" and "rock", for native Japanese speakers. Because this contrast does not exist in their first language (Flege, Takami & Mann, 1996; Goto, 1971), native Japanese speakers may engage in what Flege (1987, 1995) described as "equivalence classification" when producing these phonemes, assimilating both sounds to the Japanese /ɹ/ (phonetically, Japanese /ɹ/ is closer to the English /l/ than the English /r/; Best & Strange, 1992). As a result, Aoyama et al. (2004) found that around 20 to 25% of the time, English listeners confused Japanese speakers' /r/ and /l/ productions with each other (in a forced-choice task where they could choose among 8 consonants; confusion rates were even higher for non-native children's productions, with /r/ being misidentified as /l/ 42.8% of the time).

Comparable results have been found for other language pairs. For example, Mandarin speakers' productions of the English /θ/-/s/ and /ð/-/z/ contrasts are often misperceived by native English speakers (Hanulíková & Weber, 2010; Picard, 2002; Rau, Chang & Tarone, 2009; Rogers & Dalby, 2005; Teasdale, 1997; Zhang & Xiao, 2014). Rogers and Dalby (2005) reported that /θ/ productions are misperceived as /s/ about 30% of the time, and /ð/ is misperceived as /z/ about 20% of the time. Similarly, German speakers have difficulties contrasting the English vowels /ɛ/ and /ae/. When native English listeners were asked to categorize the words "bat" and "bet" recorded by inexperienced and experienced native German speakers, they mistook "bet" for "bat" 13% to 23% of the time, and "bat" for "bet" 34% to 49% of the time (Bohn & Flege,

1992). Strikingly, a study looking at relatively inexperienced Spanish speakers' productions of Dutch vowels found that the Dutch vowel /œy/ was misperceived as /ɔu/ as often as 89% of the time (as reported by two phoneticians, Burgos et al., 2014). Thus, this vowel was nearly always perceived as a different one by native speakers. Such productions, which fall unambiguously into an unintended phoneme category, have been called "*bad maps*" (Sumner, 2011).

In Table 1, we summarize findings regarding how native listeners perceive non-native productions. This review suggests that non-native speakers' productions are sometimes more than ambiguous, with misperception rates well above 50%. Instead, phonemes are perceived as fully substituted for others. The frequency of these substitutions, or bad maps, varies from speaker to speaker (depending on experience with the language, for example) and from one phoneme contrast to the next. As a result, there is no single number to summarize how often bad maps occur for speakers with an accent, but Table 1 indicates that these are reasonably frequent. Therefore, if perceptual recalibration is to be a mechanism for accented-speech accommodation, it is important to explore how it operates in such bad map scenarios – we must move beyond the "perfectly ambiguous phoneme" paradigm.

Reference	Number of speakers	L1	L2	L2 experience	Phonemes	Raters	Task	% misperceived
Aoyama et al. (2004)	16	Japanese	English	Low to high	/l, ɹ, w/	12 native English speakers	8AFC	<5% to 70%
Best et al. (2015)	8	London and Yorkshire English (regional accents of English)			/æɪ, əʉ, ɐ, ʊ, æ, ɐː/	80 native Australian speakers	AFC with multiple options (unclear how many)	37% to 67%

Bion et al., (2008)	5	California English (regional accent of English)			/i, I, ɛ, æ/	11 North Carolina English speakers and 11 Welsh English speakers	4AFC	<5% to 32%
Bion et al., (2008)	15	Dutch (5), Spanish (5), Portuguese (5)	English	7 years of English education	/i, I, ɛ, æ/	11 North Carolina English speakers and 11 Welsh English speakers	4AFC	<5% to 72%
Bohn & Flege (1992)	10	German	English	Low to high	/i, I, ɛ, æ/	3 native English speakers	7AFC	<5% to 50%
Burgos et al. (2014)	23	Spanish	Dutch	Low to high	14 vowels, 21 consonants	2 native Dutch phoneticians	Phonetic transcriptions	<5% to 89%
Cebrian (2007)	30	Catalan	English		/I, i, ɛ, e/	8 native English speakers	6AFC	<5% to 29%
Cutler et al., (2005)	1	American English (regional accent of English)			/i, I, e I,ɛ.æ. ɔ, oʊ, u, ɚ, ɑ, ʌ, ʊ,aI ,	10 native Australian speaker	15AFC	<5% to 82.5%
Fabra & Romero (2012)	27	Catalan	English	Low to high	/i, I, ɛ,æ,ɑ, ʌ, ʊ, u/	5 native English speakers	6AFC	30% to 57%
Flege, Bohn & Jang (1996)	80	German, Spanish, Korean, Mandarin	English	Low to high	/i, I, ɛ, æ/	3 native English speakers	7AFC	<5% to 82%
Hanulíková & Weber (2010)	74	Dutch, German	English	High	/ θ , t, s, f/	3 native speakers (1 trained phonetician)	Phonetic transcriptions	40% to 50%
Jia et al. (2006)	169	Mandarin	English	Low to high	/i, I, e, ɛ, æ, u, ʊ, o, ɔ, ʌ, ɑ /	5 native English speakers	12AFC	5% to 55%
Levy & Law (2010)	27	English	French	Low to high	/i, y, u, ɛ, œ, o, a/	9 native French speakers	7AFC	10% to 45%
Roger & Dalby (2005)	8	Mandarin	English	Low to high	18 vowels and 26 consonants	45 native English speakers	2AFC & sentence transcription	<5% to 60%

Zhang & Xiao (2014)	32	Mandarin	English	9 years of English education	/ θ, ð, s, z, v, ʃ, ʒ /	2 ESL teachers	Pronunciation judgments	17% to 50%

Table 1. Non-exhaustive list of accented phoneme misperception studies. Most involved small groups of native English listeners asked to report what they heard.

Previous research has shown that listeners can learn, through exposure, to accept a "bad map" production as a correct form for a word (e.g., Cooper & Bradlow, 2018; Cooper et al., 2023; Maye et al., 2008; Samuel & Larazza, 2015; Weber et al., 2014). For example, Maye et al. (2008), created a novel English accent in which front vowels in words like "witch" were lowered, making the pronunciation similar to "wetch". Lexical endorsement of such new, accented forms, increased significantly after exposure. However, very few studies have investigated how bad maps affect lexically-driven perceptual recalibration directly (Babel et al., 2019; Sumner, 2011; see also the unpublished dissertation Weatherholtz, 2015).

Sumner (2011) presented listeners with English words beginning with /p/ or /b/ (e.g., "paint") recorded by a native French speaker. While both English and French have the /p/-/b/ contrast, their acoustic features vary. In particular, the voice onset time (VOT) for French /p/ falls within the VOT range for English /b/. As a result, the word "paint" produced by a native French speaker may be perceived as "baint" instead (i.e., a bad map). In addition to the speaker's accent, VOTs were also manipulated artificially to fall within a more or less bad map range, depending on condition. The words were presented together with their spellings (in this example, "paint" would be shown on a screen), so that listeners could identify them despite the heavily distorted sounds (much like subtitles, which have been shown to help with accent accommodation, e.g., Cooper & Bradlow, 2016; Mitterer & McQueen, 2009). Despite the lexical support provided by

the spellings, there was no evidence of perceptual recalibration when the phonemes to be recalibrated always clearly fell into the unintended category. This finding suggests that bad maps may be hard for the perceptual system to recover from. However, Sumner's results do not inform us about situations where orthographic information is not available, which is usually the case when listening to (native and non-native) speech. In addition, the critically mispronounced segments always occurred word-initially, and this may block perceptual recalibration even for ambiguous sounds (e.g., Jesse & McQueen, 2011).

More recently, Babel et al. (2019) investigated how distortions that are not perfectly ambiguous, such as bad maps, affect perceptual recalibration. They used a lexical decision task (the most common exposure task in perceptual recalibration studies) to expose listeners to altered /s/ words (the /s/ was always word-medial, e.g., in "carousel"), as well as control words that contained prototypical /ʃ/ sounds and filler words that contained no sibilants. There were four conditions that differed in how altered the "s" items were: relatively prototypical (70% /s/ identification), ambiguous (50%), atypical (30%) and remapped (i.e., bad maps, where the wrong phoneme replaced the intended one, as in "caroushel"). After this exposure, listeners categorized sounds on an /s/-/ʃ/ continuum embedded in minimal pair items recorded by the same speaker (e.g., "sack" and "shack"). As expected, listeners exposed to ambiguous "s" in disambiguating words categorized more items on the /s/-/ʃ/ continuum as "s" – this is the typical perceptual recalibration result. In contrast, there were only nonsignificant trends towards recalibration for the atypical and bad map conditions, suggesting that listeners learn the most when the phoneme distortion is perfectly ambiguous. However, using a lexical decision task to expose listeners to the distorted sounds may not be the best way to test recalibration for atypical or bad map sounds.

In fact, Babel et al. reported that listeners perceived bad map items as nonwords 70% of the time – a much higher error rate than for the ambiguous items (about 20% error). As the authors noted, this is a potential problem because lexically-driven perceptual recalibration relies on listeners recognizing that the distorted phoneme is part of a word and using that knowledge to infer what the sound should have been (e.g., Norris et al., 2003). If listeners are not hearing the exposure items as words, the conditions do not meet the requirements for recalibration to occur.

Given that there have only been two relevant studies, and each is best viewed as preliminary, whether perceptual recalibration occurs after exposure to bad maps is unknown. The current study is intended to provide evidence to answer this question. Our approach shares features with those of Babel et al. (2019) and Sumner (2011), but with changes designed to overcome some of the methodological concerns described above. As is typically the case, we included two tasks: an exposure task that served to familiarize participants with the accented phonemes, and a test task that measured each listener's phoneme category boundary. Critically, we chose an exposure task that would strongly encourage listeners to treat all items as words despite clear pronunciation "errors" (e.g., "shuperpower" for "superpower"). As we just noted, this is important to observe recalibration effects. It is also more similar to a real-world scenario, in which listeners assume that their non-native interlocutors are producing real words, not nonwords. In addition, we manipulated the position of the critical phonemes since, as mentioned, this has been shown to affect recalibration (Jesse & McQueen, 2011; Samuel, 2016). If recalibration to bad maps occurs in the same way as it does for ambiguous sounds, we should observe similar positional effects.

Crossing this positional manipulation with the contrast between ambiguous and bad map stimuli yielded four exposure conditions: 1) ambiguous sounds occurring late in words (Late Ambiguous); 2) ambiguous sounds occurring early in words (Early Ambiguous); 3) bad map sounds occurring late in words (Late Bad Map) and 4) bad map sounds occurring early in words (Early Bad Map). We used the /ʃ/-/s/ contrast, which has been one of the most commonly used contrasts in perceptual recalibration studies and therefore a good starting place to explore new situations (i.e., the bad maps). Our experiment included only the /s/ side of this contrast to keep the sample size needed within available resources (as will become clear, even with this limit we required a very large sample). The /s/ side of the contrast was chosen because, in previous work in our lab, it has consistently produced robust effects. All of the stimuli were recorded by a male native speaker of English, and the critical phonemes were then artificially accented through editing, by mixing /s/ and /ʃ/ productions. One large group of participants completed only the test task without any prior exposure to the accented speech, providing a baseline for the four experimental conditions.

We report two versions of this experiment. A first version of the experiment was conducted in the lab. When the Covid-19 pandemic effectively shut down in-person testing, we conducted an online replication of the in-person experiment. This was not part of our original plan, but (a) we wished to follow current best-practices of replicating findings, and (b) we anticipated future experiments exploring bad maps that were likely to be online and wished to be sure that we could compare the results of the current study to any such future experiments. The decision to run the same experiment in two different settings allowed us to observe how experimental setting (in-lab vs. online) affects the observed recalibration results.

Method

Participants.

In total, 218 Stony Brook University undergraduate students participated in the in-lab version of the experiment. All participants were 18 or older and reported being native speakers of English with no known hearing problems. Of these 218 in lab participants, 140 were assigned to one of the four experimental condition (i.e., Early Ambiguous, Late Ambiguous, Early Bad Map or Late Bad Map) and 78 to the Baseline condition. In addition, 505 Stony Brook undergraduate students participated in the online version of the experiment. Of these, 198 did not report using some form of headphones and were therefore not included in the dataset. Of the 307 remaining participants, we excluded from further analyses: 2 participants who did not report having normal or corrected to normal hearing and 50 participants who did not report English to be their native language. This brought the total of online participants eligible for analyses down to 255, with 160 assigned to one of the four experimental conditions and 95 to the Baseline condition.

Following Norris et al.'s (2003) seminal work, there have been scores of studies on recalibration, including a substantial number from our lab. Drawing on this experience, the targeted number of participants for each experimental condition was 35, a slightly larger sample than in many previous recalibration experiments (e.g., in-lab: ~24 in Kraljic & Samuel, 2007, using the same contrast; ~25 in Norris et al., 2003; online: ~24 in Kleinschmidt & Jaeger, 2012). For the Baseline conditions, we aimed for a larger sample size to ensure it was a representative measure of the /s/-/ʃ/ categorization function for native English listeners. This was possible

because the baseline task was short (about 5 minutes) and could be run before participation in unrelated experiments.

Materials.

Exposure task:

In recalibration studies, an exposure task is typically used to present 15-20 words that each have a critical sound intended to drive the recalibration. The most common such task has been lexical decision. When the critical sound is ambiguous most listeners do not notice the mispronunciation. However, with "bad map" stimuli, the mispronunciation is very noticeable, causing listeners to frequently classify the exposure word as a nonword (Babel et al., 2019). Therefore, we designed an exposure task that, unlike lexical decision, did not encourage participants to be listening for nonwords. In this "Noun task", the listeners are told that they will hear a list of words, and that their task is to classify each word as either a noun, or not a noun. By encouraging listeners to assume that every item was a word, we maximize the likelihood that even items with mispronunciations will achieve lexical access, a necessary aspect of driving recalibration.

We created two experimental lists for the Noun task, each with 50 nouns and 50 non-nouns (i.e., verbs, adjectives or adverbs). The two lists were identical except for 20 critical words: In one list, all critical words contained an /s/ phoneme that occurred after the words' uniqueness point (i.e., late in the words); in the other, the /s/ occurred before the words' uniqueness point (i.e., early in the words). Note that while the nouns to non-nouns ratio was one to one within each list, it was not exactly one to one for the 20 critical items (which already had a

number of constraints on their selection). The lists can be found in Appendix A. The lists did not contain words with unambiguous /ʃ/ sounds; previous studies show these are not necessary to observe a recalibration effect (e.g., Jesse, 2021; Zheng & Samuel, 2023).

Critical words. We selected 40 words ranging in length from two to five syllables. These critical words could be nouns or "not nouns". Each contained a single instance of the critical phoneme /s/, and no instances of /ʃ/. The /s/ phoneme never occurred in a consonant cluster (aside from "rs", where 'r' is produced as a coloring of the vowel /ɚ/ or /ɝ/). Half of these critical words had the critical /s/ phoneme after the uniqueness point of the words (i.e., Late in words), and in half the /s/ occurred before the uniqueness point (i.e., Early in words). For each word the uniqueness point was defined as the point at which the word's initial sequence of phonemes was not shared with any other morphologically unrelated words; at this point, there is enough lexical information to identify the word. The uniqueness points were determined with a phonetic dictionary (the Carnegie Mellon Pronouncing Dictionary, 2014).

Filler words. We selected 82 words ranging from two to four syllables in length. Like the critical words, filler words included nouns, verbs, adjectives and adverbs, but they contained no instances of /s/ or /ʃ/. Two of the filler words ("recover" and "royal") which were selected as "non-nouns" were also recorded in a noun form ("recovery" and "royalty") to keep the noun/not-noun ratio equal in the two experimental lists.

Stimulus construction. All words were recorded by a male American English speaker. Critical words were recorded twice: once in their canonical /s/ form and once with /ʃ/ substituting for the critical /s/ (e.g., "superpower" and "shuperpower"; "malpractice" and "malpractish"). We

used PRAAT (Boersma & Weenink, 2018) to create a 21-step continuum for each word, with the critical sound ranging from /s/ to /ʃ/. To do so, we used a PRAAT script written by Mitterer (n.d.) that matches two recordings for pitch contour and then gradually merges them in 5% increments, such that the resulting continuum steps are weighted composites of the original two items. So, for example, the second step of a continuum between "superpower" and "shuperpower" would be a merged form of the two recordings in which the fricative is 95% /s/ and 5% /ʃ/ (note that the rest of the word is also the result of merging), the third step is 90% /s/ and 10% /ʃ/, and so on . Based on ratings from eight phonetically untrained raters, we selected a single step for each word to serve as the "ambiguous"' version of that word (see Appendix D). For the bad map version of the words, we simply used the original /ʃ/ recordings (e.g., *shuperpower*). Four lists were created, one for each experimental condition: Early Ambiguous, Late Ambiguous, Early Bad Map and Late Bad Map. The lists varied in 1) whether the critical /s/ phoneme was ambiguous or a bad map and 2) whether the critical /s/ phoneme occurred before or after the words' uniqueness point, that is, whether it occurred early or late in the words. See Table 2 for a summary and examples.

		Critical phoneme manipulation	
		Ambiguous	Bad Map
Critical phoneme position	Early	[?ssh]uperpower	[sh]uperpower
	Late	malpracti[?ssh]	malpracti[sh]

Table 2. Summary of the four conditions, with example stimuli.

Task 2: Test (phoneme categorization task).

The same male American English speaker who produced the Noun task stimuli recorded the nonwords “asee” (/ɑsi/) and “ashee” (/ɑʃi/) with stress on the second syllable. The two were mixed with PRAAT to create 21 mixtures ranging from very s-like to very sh-like. We selected seven steps centered on the most ambiguous one, based on ratings from native English speakers (see Appendix D for more details). The steps chosen ranged from relatively /s/-like to relatively /ʃ/-like, with ambiguous points in between.

Procedure.

Participants were randomly assigned to one of the four experimental conditions. For the in-lab version, up to three participants were tested simultaneously in a soundproof booth. People in the online version did the task individually and were asked to do so in a quiet setting. During the Noun task, participants were instructed to respond “Noun” or “Not a Noun” for each word presented to them, using a labelled button pad in-lab, and the letters A and L on the keyboard in the online version. For clarity, noun and “non-noun” examples were provided during the instructions. For the few items that could fit in both categories, participants were told to use their best judgment (note that performance on this task is not crucial to the experiment – it simply serves to expose participants to the accent).

Participants were not told that some of the words would have ambiguous or bad map sounds. The words were presented in a random order and participants had a maximum of 5 seconds to provide an answer. Directly after this exposure task, participants performed the phoneme categorization task. The seven asee-ashee steps were presented in a random order, with

ten repetitions each, totaling 70 trials. For each item, participants were instructed to indicate whether they heard /s/ or /ʃ/ using labelled buttons in-lab, and the keys A and L in the online version. In addition, in each version, a group of participants assigned to the Baseline condition performed the phoneme categorization task without any prior expose task.

The procedure was similar for the in-lab and online versions of the experiment. One difference was that online participants received all instructions in writing instead of orally, and the tasks started once they left the instruction pages via a button click. In addition, both the Noun task and the phoneme categorization task were preceded by short tutorials to familiarize participants with the tasks (this was designed as a substitute for participants being able to ask questions when experiments are done in the lab). Finally, once participants had completed both tasks, they were asked to answer a short questionnaire asking about their language background and headphone use before exiting the experiment (see Appendix B).

Results

Of the 218 in-lab and 255 online participants, 29 in-lab participants and 37 online participants were excluded based on the following criteria: 1) less than 50% accuracy on the noun items and/or the non-noun items during the Noun task (in lab*:* n = 8*;* online*:* n = 3); and/or 2) less than a 50% change in S responses between the first and last step of the /s/-/ʃ/ continuum during the phoneme categorization task (in lab: n = 21; online: n = 34). If identification of the two endpoints does not differ very much, it indicates that a participant was unwilling or unable to do the required identification. After excluding these participants, the sample sizes for each condition were as follows: 30 in-lab and 36 online participants in the Early Ambiguous

condition; 32 in-lab and 36 online in the Late Ambiguous condition; 28 in-lab and 35 online in the Early Bad map condition; 31 in-lab and 35 online in the Late Bad map condition; and 68 in-lab and 76 online in the Baseline condition.

All results were analyzed using generalized linear-mixed effects models (lme4 package, Bates, Maechler, Bolker, & Walker, 2015; in R, version 4.0.2) with a logistic linking function to account for the categorical nature of the dependent variables (Accuracy in the Noun task, and Phoneme identity in the Phoneme Categorization task). If they did not improve model fit, interactions between predictor variables were not included. Following Barr et al. (2013), we used maximal random-effect structures when building models (including all within-subject variables and their interactions). In cases where models did not converge, they were simplified by taking out random interaction effects first, followed by simple random effects when necessary. Models were compared using likelihood ratio tests. The full statistical outputs for all analyses can be found in Appendix F.

Noun task.

The main variable of interest for the Noun task was accuracy (1 = accurate, 0 = inaccurate). The predictor variables were Condition (Early Ambiguous, Late Ambiguous, Early Bad Map, Late Bad Map; Late Ambiguous was chosen as the reference level as this is the condition most often studied in the literature), Grammatical Category (i.e., Noun or Not-noun, with Noun as the reference level), Item Type (i.e., Critical item versus Filler item, with Filler as the reference level) and Trial. Note that the Condition factor pertains to potential effects on the identification test that followed the Noun task; no effect of this factor would be expected on the

Noun task itself. The models included random intercepts for Items and for Participants, as well as random slopes for Item Type and Grammatical Category over Participants.

Both in-lab and online, average accuracy was high (especially given the grammatical category ambiguity of some items), averaging 85.9% and 87.8% respectively (see Appendix C for a plot of the individual conditions). There were no significant differences among the four conditions in-lab (χ^2= 3.95, df = 3, p = 0.27) or online (χ^2= 2.01, df = 3, p = 0.54), reflecting the random assignment of participants to the groups. On average, participants were slightly more accurate on filler trials in both experimental settings, suggesting that the distortions in the critical items impacted processing to some extent. This difference was not significant in-lab ($M_{critical}$ = 82.2%, M_{filler} = 86.8; p= 0.21), and was marginal online ($M_{critical}$ = 84.4%, M_{filler} = 88.7%; p = 0.063).

While there was no difference in accuracy between nouns and non-nouns for in-lab participants (M_{nouns} = 82.7%, $M_{non\text{-}nouns}$ = 89.1%; p = 0.62), online participants were better at identifying non-nouns (M =92.8%) compared to nouns (M = 82.9%; b = 0.6956, SE = 0.3350, *Wald's z* = 2.08., p = 0.038). In fact, a cross experiment analysis showed that online participants were slightly more accurate than the in-lab ones overall on the Noun task (*Min-lab* = 86% vs. *Online* = 88% - b = 0.2543, SE = 0.0869, b = 2.93, p = 0.0034), but despite this difference between the two versions of the experiment, there was no interaction between experimental setting and condition (χ^2= 1.44, df = 3, p = 0.7). This suggests that participants did not treat the critical items differently across settings.

Phoneme Categorization task.

We assessed participants' categorization of sounds on the /s/-/ʃ/ continuum to examine whether prior experience mattered. The key comparison is categorization in each experimental condition (i.e., Late Ambiguous, Early Ambiguous, Late Bad Map, Early Bad Map) versus the Baseline condition (where participants received no prior exposure). If participants in an experimental condition were affected by exposure to the "accented" /s/ sounds during the Noun task, we should find a significant difference between that condition and the Baseline in terms of the number of tokens categorized as "s".

The dependent variable was "S Answer" (1 = S response, 0 = SH response). Models included the predictor variables Condition, Step (centered on zero and ranging from -0.5 to 0.5) and Trial. Condition had five levels corresponding to our four experimental conditions plus the Baseline condition. The latter was set as the reference level, meaning that performance in each experimental condition was compared to performance for the Baseline group. A significant difference between the two would indicate a difference in the percentage of S responses on the categorization task, i.e., a boundary shift between the /s/ and /ʃ/ phoneme categories. We did not include the interaction of Condition and Step as it did not improve model fit (χ^2= 5.36, df = 4, p = 0.25). Finally, models included random intercepts for Participant and random slopes for Step over Participant. Only responses to the five middle steps of the continuum were included in the analysis (following similar procedures in, e.g., Samuel, 2016), because shifts near the end points of the continuum are typically limited by floor/ceiling effects. As expected, there was an overall significant effect of Step both in lab and online: Participants made fewer S responses as the steps became more /ʃ/ like (in lab: b = -7.333, SE = 0.227, *Wald's* z = -32.26, p <0.001; online: b= -6.835, SE = 0.199, *Wald's* z = -34.29, p < 0.001).

There are recent indications in the literature that perceptual recalibration effects may attenuate as the phoneme categorization task unfolds (Liu & Jaeger, 2018, 2019). If such attenuation occurs, then the data from later trials will dilute the apparent size of any shifts. To determine if this occurred in the current study, we analyzed the results separately for the first five repetitions for each step and the last five repetitions, using the same mixed model analyses. In these analyses, we focused on the "standard" recalibration condition, the Late Ambiguous case, which we know from a large body of prior work should show a reliable shift. Indeed, for the data collected during the first five presentations, we find a reliable shift for both the in-lab version (an 11.0% shift, $b = 0.970$, $SE = 0.269$, *Wald's* $z = 3.61$, $p < 0.001$), and the online version (a 9.7 % shift, $b = 0.856$, $SE = 0.276$, *Wald's* $z = 3.09$, $p = 0.002$). In contrast, for the data collected during the last five presentations, the effect disappeared completely, both in the lab (a shift of -0.01%, $p = 0.90$) and online (a shift of 0.9%, $p = 0.718$). Given our confirmation of the pattern reported by Liu and Jaeger, the analyses we report in the text are only based on data from the first five presentations. Figure 1 presents these results for the Ambiguous conditions, and Figure 2 shows the corresponding results for the Bad Map cases. We will discuss the in-lab results first (i.e., those that were collected under conditions like those in most of the literature), and then consider the online results. The results for the full data set, and for the last five presentations, can be found in Appendix E.

In-Lab Testing: As we just noted, we find the expected recalibration effect in the classic recalibration testing conditions: ambiguous sounds occurring late in words, presented in a lab setting. What of early ambiguous sounds in a lab setting? Aligning with the past literature (e.g., Jesse & McQueen, 2011), we find a small shift for the Early Ambiguous condition (4.8%) (the comparable shift for the Late Ambiguous case was 11.0%), with the Early Ambiguous shift only trending in the expected direction ($b = 0.499$, $SE = 0.276$, *Wald's* $z = 1.81$, $p = 0.070$).

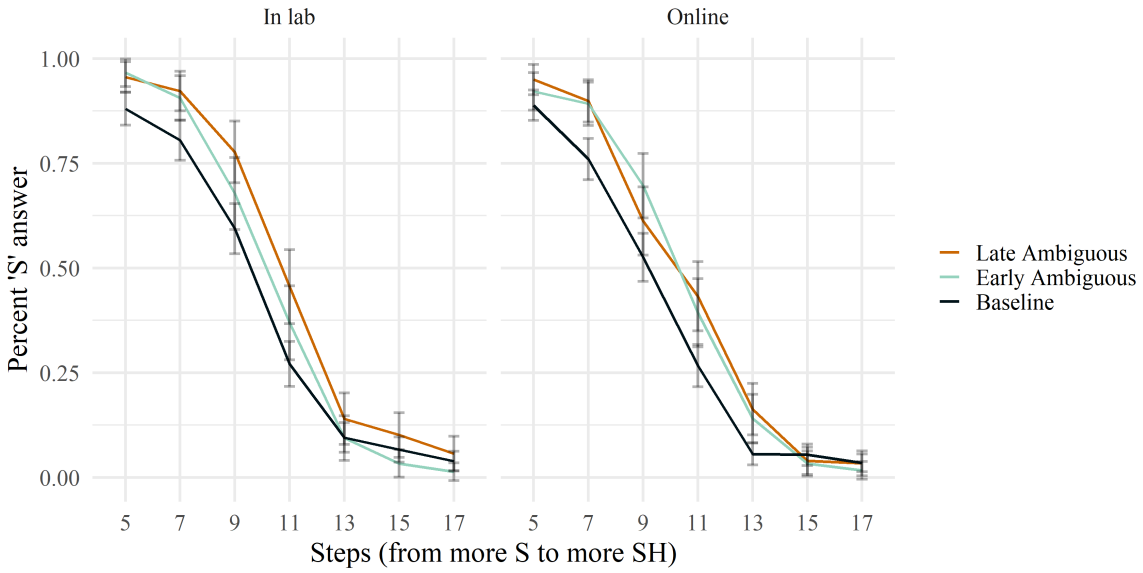


Figure 1. Categorization functions for the Baseline (in black), Late Ambiguous (orange) and Early Ambiguous (light blue) conditions for the first five repetitions of each step. Results obtained in the lab are pictured on the left and results obtained online are pictured on the right. In lab, only the Late Ambiguous condition produced a significant shift, whereas online the shifts were significant for both the Late and Early Ambiguous conditions. Error bars represent the standard error of the mean.

Having replicated the standard findings from the recalibration literature, we turn to our primary question: Does recalibration manifest in the same way for bad maps as it does for ambiguous sounds? If so, then the Late Bad Map condition, in the lab, should show robust recalibration. However, as the left panel of Figure 2 shows, we find almost no recalibration shift for this condition (a 2.0% shift, $b = 0.2203$, $SE = 0.2683$, *Wald's* $z = 0.82$, $p = 0.411$), indicating that a clear mispronunciation occurring late in a word may block recalibration. In contrast, bad maps occurring *early* in words did produce a significant recalibration effect (a 6.1% shift, $b = 0.631$, $SE = 0.283$, *Wald's* $z = 2.23$, $p = 0.026$). Thus, the effect of Position may be different for the Bad Map critical segments than for the Ambiguous critical segments. More formally, there appears to be an interaction between how distorted a phoneme is (i.e., Ambiguous vs. Bad Map) and its position in a carrier word (i.e., Early vs. Late).

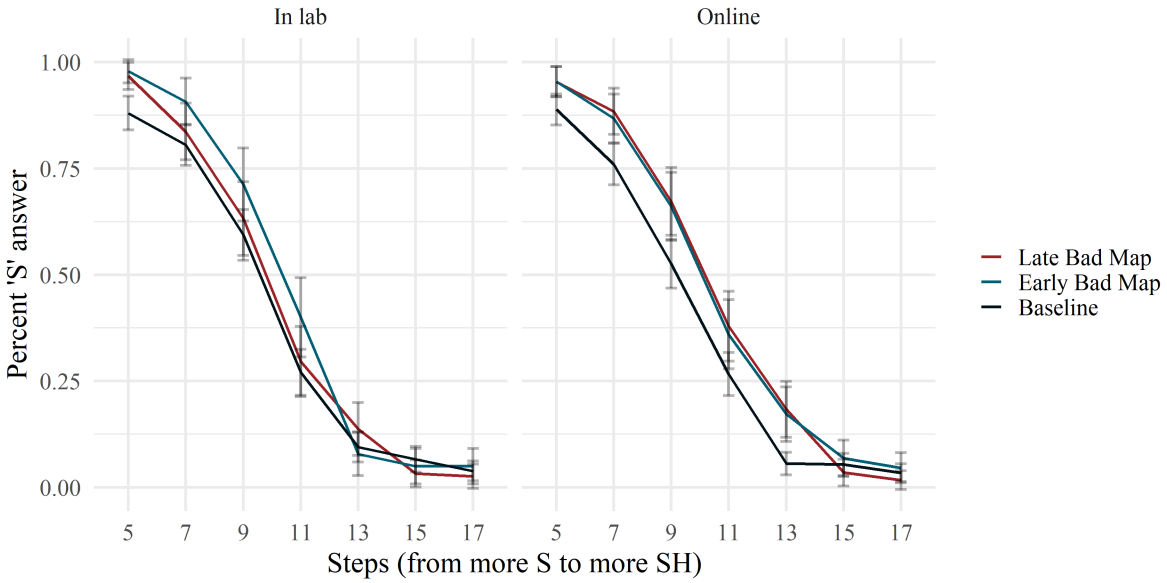


Figure 2. Categorization functions for the Baseline (in black), Late Bad Map (red) and Early Bad Map (dark blue) conditions for the first five repetitions of each step. Results obtained in the lab are pictured on the left and results obtained online are shown on the right. In the lab, only the Early Bad Map condition produced a significant shift, whereas online the shifts were significant for both the Late and Early Ambiguous conditions. Error bars represent the standard error of the mean.

We conducted additional analyses using a generalized mixed effect model that included the factors Step, Trial, Pronunciation, Position and the interaction between Pronunciation and Position. Pronunciation had three levels based on what listeners heard during the exposure phase: 1) ambiguous sounds (i.e., Ambiguous); 2) bad map sounds (i.e., Bad Map) or 3) no critical sounds (i.e., None; the reference level). Similarly, Position included levels Late versus Early for the critical items and "None" to account for the control items (here Late was chosen as the reference level because we were interested in the comparison between Late and Early). The model included random intercepts for Participants. There was a marginal interaction between Pronunciation and Position (b = -0.882, SE = 0.453, *Wald's* z = -1.95, p = 0.051), suggesting that result patterns for each pronunciation type (Ambiguous or Bad Map) may be affected differently by position in the carrier words.

Online Testing: Do we observe the same result patterns online? The answer is both yes and not quite (see the right panels of Figure 1 and 2). As in the lab, the online experiment yielded reliable recalibration shifts for the Late Ambiguous (a 9.7 % shift, b = 0.856, SE = 0.276, *Wald's* z = 3.09, p = 0.002) and the Early Bad Map conditions (b = 0.771, SE = 0.279, *Wald's* z = 2.77, p = 0.006). Where the recalibration shift was only trending in the expected direction for the Early Ambiguous case in the lab, here we observe a significant effect (b = 0.825, SE = 0.281, *Wald's* z = 2.94, p = 0.003), with a shift size equivalent to that of the Late Ambiguous condition (a 9.7% shift in both cases – right panel of Figure 1). For these three conditions, results across experimental settings are rather consistent. However, in contrast with the in-lab results, online participants showed a relatively strong recalibration effect in the Late Bad Map condition (b = 0.829, SE = 0.279, *Wald's* z = 2.97, p = 0.003).

Table 3 summarizes these results. For the "classic" recalibration stimuli – ambiguous segments placed late in carrier words – recalibration holds for both in-lab and online testing. However, the nuances observed for other conditions in the lab, mainly the lack of effect for the Late Bad Map condition, seem to be largely lost online. In fact, in this setting, the sizes of the recalibration shifts are strikingly similar across all conditions. Interestingly, when the data from both versions of the experiment are combined, providing us with unusually large sample sizes for such experiments, and analyzed with the same model, we see significant recalibration effects for all conditions (see Appendix E for the statistical analyses), with the largest shift from the Baseline for the Late Ambiguous condition (10.3% shift), followed by the Early Bad Map (7.8% shift), Early Ambiguous (7.4% shift) and Late Bad Map (6.2% shift) conditions. One way to view this ordering may be as a reflection of the likelihood, across stimuli, of there being an optimal combination of lexical activation and acoustic cues to drive recalibration - we discuss this further below.

	% ‘S’ answer	
Condition	In Lab	Online
Baseline	36.8	33.2
Late Ambiguous	47.8	42.9
	(Δ11.0)	(Δ9.7)
Early Ambiguous	41.6	42.9
	(Δ4.8)	(Δ9.7)
Late Bad Map	38.8	43.1
	(Δ2.0)	(Δ9.9)
Early Bad Map	42.9	42.5
	(Δ6.1)	(Δ9.3)

Table 3. Percentages of S answers for all four experimental conditions and the Baseline conditions for the in-lab and online experiments, for the middle five steps of the continuum. Differences from the Baseline are shown in parentheses.

To further explore how experimental setting may have influenced our results, we developed a measure that is designed to norm the experimental conditions across the different settings: We subtracted the average percentage of S answers for each continuum step for the Baseline (either the in-lab, or the online baseline, as appropriate) from the corresponding numbers for each of the experimental conditions. This gives us a difference score that represents the size of the recalibration shift at each step, accounting for any differences in the baselines in the different settings.

Figure 3 shows the resulting functions, illustrating the recalibration pattern in the four cases (i.e., the crossing of Early/Late with Ambiguous/Bad Map conditions). The uniformity of the online setting results can easily be seen – the dark curves are similar in all four panels. In contrast, there is clearly more variation in the lab setting. Comparing the two, the lighter curves are similar to the darker ones for the Late Ambiguous case (the strongest effects) and for the Early Bad Map case (a moderate effect). The biggest difference is evident in the Late Bad Map case: There is virtually no effect in the lab, but the shift is comparable to that seen in the other conditions tested online, reflected in the separation between the light and dark curves in this panel.

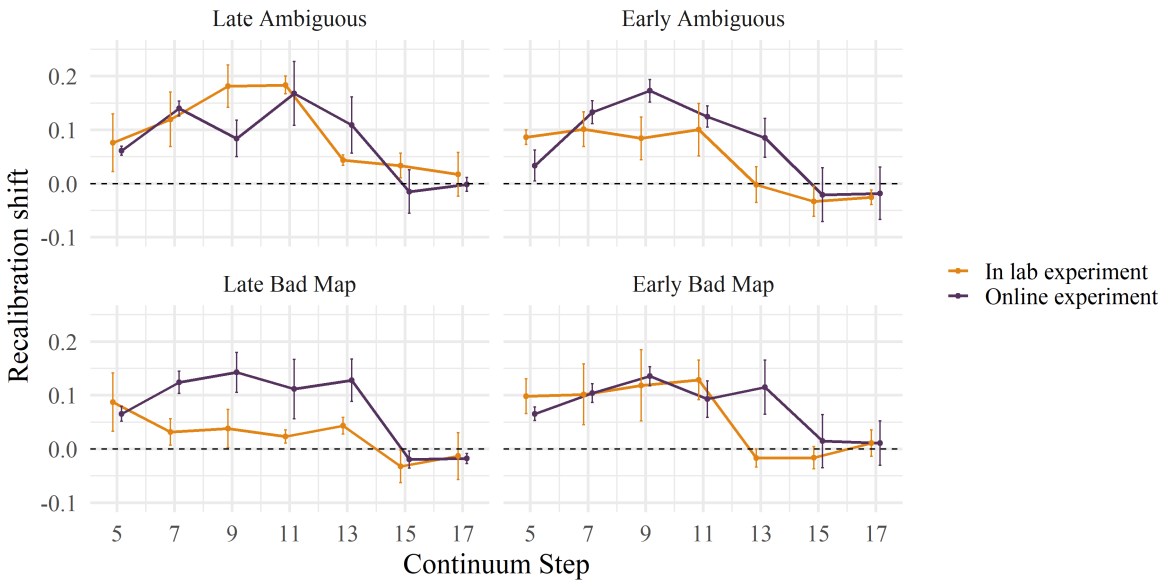


Figure 3. Recalibration shifts for each step of the continuum in the in lab (orange) and online (purple) versions of the experiment. The dotted line represents 0, or no difference between the experimental condition and the baseline. A positive shift means that experimental participants categorized more sounds as S than baseline participants (i.e., the expected recalibration effect). Error bars represent the standard error of the mean.

We used a linear mixed-effect model to formally assess whether these difference scores differed online versus in-lab. The model included the fixed factors Condition (reference level: Late Ambiguous), Step (centered on 0), Experiment (in-lab vs. online) and the interaction of Condition and Experiment. It also included random intercepts for Participants. For the Early Ambiguous, Late Ambiguous, and Early Bad Map conditions, there was no interaction with test setting. For the Late Bad Map condition, the apparent interaction seen in Figure 3 between Condition and Experiment did not reach significance ($b = 0.092$, $SE = 0.052$, $df = 255$, $t = 1.76$, $p = 0.080$). Thus, although it appears that online testing may miss some nuances found in the lab, we cannot say with certainty that this is the case.

Discussion

Lexically-driven perceptual recalibration is often assumed to underlie accommodation to accented speech, despite the lack of direct evidence for such a link. In assessing this assumption, an important feature of the perceptual recalibration literature is that it has mostly focused on perfectly ambiguous phonemes embedded in clear words. If the segmental distortions resulting from accents are consistently ambiguous, then the connection would have at least face validity. Surprisingly, we were not able to find any systematic reviews that looked at whether accent-based distortions do in fact consistently yield such ambiguity. Therefore, we reviewed a large number of papers that examined accented speech, and found that the literature does not support the assumption that accented segments are consistently ambiguous. Rather, as Table 1 shows, there is a mix of possible outcomes, including a substantial likelihood that a segment will be

perceived as another, rather than as an ambiguous compromise. Following Sumner (2011), we have called such substitutions “bad maps”.

Given these results, for recalibration to be established as a convincing mechanism driving accommodation to accented speech, the standard paradigm must be scaled up to a broader range of scenarios that are more representative of listeners’ experience. This was the main goal of the current study. Toward this end, we examined both the often-studied ambiguous critical sounds, and the little-studied bad maps. We did so for critical sounds that occurred before, or after, enough lexical information was available to uniquely identify their carrier words. Importantly, we used an exposure task that was designed to work for both ambiguous sounds and for bad maps. These changes allowed us to address our core theoretical question: Do we observe recalibration after exposure to bad maps in the same way that we do after exposure to ambiguous sounds?

Our findings suggest that the answer is yes, but with some important caveats. We did indeed find recalibration effects for bad maps. Interestingly, when these effects occurred, they were comparable to those for ambiguous sounds. That is, listeners did not show numerically larger boundary shifts (which could have been expected given the greater distortion of the phoneme), and the shift was not concentrated towards the endpoints of the continuum (i.e., listeners did not learn to categorize /ʃ/ sounds as /s/ -- they just accepted more ambiguous sounds as /s/). These results significantly extend Babel et al.’s (2019) report of a (nonsignificant) tendency for bad maps to support recalibration. As those authors noted, their use of a lexical decision exposure task may have been responsible for the lack of a significant recalibration effect

because most participants failed to recognize the bad map words as lexical items. Lexicallydriven recalibration relies heavily on listeners recognizing the distorted phoneme's carrier items as words (e.g., Norris et al., 2003). Our use of a different exposure regime -- the Noun task -- was specifically designed to encourage the participants to treat all items as words. Taken together, the results of both studies suggest that recalibration can occur after exposure to bad maps.

That said, our results also highlight that optimal recalibration may require certain conditions. As suggested by previous research, the position of the critical phoneme in its carrier word is one such factor modulating the success of recalibration (e.g., Jesse & McQueen, 2011; McAuliffe & Babel, 2016; Samuel, 2016). That is, recalibration effects for ambiguous sounds are larger when enough lexical information is available to disambiguate the critical phoneme when it is heard. Our findings are consistent with this, as we found numerically smaller recalibration effects for ambiguous sounds occurring early in words compared to those occurring later. But surprisingly, this was reversed for bad map conditions, with only bad map sounds occurring early in words leading to significant boundary shifts in both versions of our experiment.

That recalibration occurs even when the disambiguating information arrives after the ambiguous or bad map phonemes adds to a growing body of evidence that the speech perception system is able to maintain uncertainty about the signal and to delay lexical commitment (e.g., Burchill et al., 2018; Connine et al., 1991; Connine et al., 1994; Kapnoula et al., 2017; Kapnoula et al., 2021; McMurray et al., 2009; Samuel, 1981; Samuel, 1991; Samuel, 2016; Szostak & Pitt, 2013). This includes recent evidence from neuroimaging studies showing that listeners preserve

and revisit at least some fine-grained acoustic and phonetic details over long timescales (i.e., several hundred milliseconds to seconds after phoneme onset; Blanco-Elorrieta et al., 2021; Gwilliams et al., 2018). Importantly, this may be the case even when uncertainty about phoneme identity is low, as in the bad map case (e.g., Gwilliams et al., 2018).

For example, McMurray et al. (2009) tested listeners' ability to recover from "lexical garden paths" in an eye-tracking experiment. A lexical garden path occurs when a word-initial ambiguous sound originally makes the identity of the word consistent with at least two interpretations (e.g., a sound ambiguous between /b/ and /p/ in "barricade" will initially make the word ambiguous with "parakeet"). In McMurray et al.'s eye-tracking task, listeners heard a target word and were to select its corresponding image amongst a set of four. In each set one image matched the target (an image of a "barricade" in our example) and one was a "competitor" (an image of a "parakeet"). Before the identity of the word became clear, listeners looked at both the target and competitor an equivalent amount of time. But by the last syllable, they fixated the target words (understanding, in our example, that the intended word was "barricade" with an ambiguous /b/ sound). McMurray et al. (2009) measured how quickly listeners were able to correct their interpretation of /b/-initial words and fixate their gaze on the correct object. This recovery time was faster when the first phoneme was ambiguous (e.g., between /p/ and /b/), rather than a bad map (i.e., clearly /p/-like). But interestingly, bad maps (e.g., "parricade") rarely fully blocked recovery, suggesting that the speech perception system remained flexible even when the bottom-up information strongly favored one interpretation for a phoneme. This aligns also with previous research showing that, as long as the mispronounced initial phoneme deviates by no more than two phonetic features from the intended sound, the lexical representation of the

intended word will be activated (Connine et al., 1993). This was the case for our bad map words, which deviated from their base words by only one phonetic feature.

Although our finding that early bad maps can lead to recalibration fits within the previous literature, it is surprising that it was found more robustly across experiments than recalibration for late bad maps. As mentioned, fine-grained acoustic details may remain available to the speech perception system during processing (e.g., Blanco-Elorrieta et al., 2021; Gwilliams et al., 2018), but it is not clear how precise this information is. We can only speculate here, but some of the information may fade, leading listener to "retroactively" perceive an early bad map "sh" sound as more ambiguous than it was, leading to boundary shifts more similar to those for ambiguous phonemes. There is also evidence that lexical activation may rely on different heuristics at the beginnings and endings of words. As mentioned, there is now some consensus that, at the beginnings of words, lexical hypotheses are activated in proportion to the bottom-up acoustic evidence, with uncertainty about phoneme interpretations modulating the strength of lexical activation of different words (e.g., Connine et al., 1994; Marslen-Wilson & Welsh, 1978, McMurray et al., 2009; Samuel, 1981). But by the ending of a word, the remaining acoustic evidence may act more like a switch, either activating or deactivating a lexical item (Gwilliams et al., 2017). In other words, it may be that when a clearly mispronounced sound occurs late in words, discrediting the previous information in favor of one interpretation, it leads to inhibition of the activated lexical item and therefore less recalibration (which relies on lexical activation, e.g., Norris et al., 2003).

Our results speak to two additional methodological issues. First, although not part of our original hypotheses, but consistent with recent findings (Liu & Jaeger, 2018, 2019), we found that the first few passes in the phoneme categorization task provide the most accurate measurement of recalibration. As Liu and Jaeger have suggested, after this point, exposure to the test stimuli themselves can produce recovery from recalibration. We observed this pattern across experimental settings (i.e., both in-lab and online) and for both ambiguous and bad map conditions. Future (and perhaps past) recalibration studies should take this into consideration.

Second, in response to the COVID-19 pandemic, we included a full online replication of our in-lab experiment, providing a direct comparison of the results obtained in-lab and online (with the same set of stimuli and procedures). In both settings, we found the classic recalibration effect for late ambiguous /s/ sounds, adding to a set of web-based studies showing that recalibration effects can be successfully observed online (e.g., Kleinschmidt & Jaeger, 2012; Kleinschmidt & Jaeger, 2015). However, there were some apparent differences in the pattern of results across the two settings. Overall boundary shifts were numerically larger online, and differences among the four conditions (Ambiguous vs. Bad Map x Early vs. Late) were largely lost. This led us to conduct the exploratory analyses summarized in Figure 3. Although we found no significant differences, it is clear that, had we conducted only one of our two experiments, our conclusions regarding recalibration to bad maps would have been slightly different. This suggests that while recalibration can be replicated online, there may be less sensitivity to the effects of modulating factors. Very few online participants are likely to have the professional quality headphones used by in-lab participants. Similarly, online participants were not in sound-shielded chambers while completing the experiment. The reduced quality/control of the listening

conditions could account for there being less sensitivity to the experimental manipulations, though it is important to keep in mind that the overall strength of the recalibration was not reduced online. There is an argument to be made that the online results are more representative of the "real world", what the speech perception system actually must deal with. In any event, our findings suggest that one should not assume that online and in-lab tests will produce the same results in detail, even if the two data sets are broadly comparable. This is an important consideration as the field moves towards an increased use of online testing.

We began this project with one fundamental question that led to two more specific ones. The general question was whether lexically driven recalibration is a primary mechanism for accommodating to accented speech. The more specific questions followed from the primary one: Are the segments in accented speech similar to the perfectly ambiguous sounds used in virtually all recalibration studies? And, if they are not – if they are instead often unambiguous substitutions – would recalibration work with such bad maps?

Our review of the literature on the perception of non-native production highlighted a potential problem: In many cases, accented speech produces substitutions, not ambiguous "compromises" between the intended and alternative sounds. Thus, we compared the ability of bad map tokens to ambiguous ones, in driving recalibration. We found that recalibration is most robust when a perfectly ambiguous sound occurs at a point when enough lexical information is available to correctly identify it – the classic conditions for recalibration. When the deviation of the critically accented phoneme is extreme, or when there is not sufficient lexical activation to resolve it, boundary shifts occur, but less reliably across experiments (Babel et al., 2019; Clarke-

Davidson, Luce & Sawusch, 2008; Jesse & McQueen, 2011; McAuliffe & Babel, 2016; Samuel, 2016; Sharenborg & Janse, 2013). Note that despite these limitations on recalibration, our results indicate that the speech perception system is flexible enough to recover and learn from non-optimal scenarios (e.g., bad map sounds, and ones occurring before disambiguating information), just not as well or consistently as it can under more optimal conditions. To the extent that this occurs, perceptual recalibration could be a possible mechanism for accommodation to non-native speech. However, we believe that the limitations that we have observed, and the fact that only a few studies so far have aimed to scale the paradigm up to a broader range of scenarios, call for caution in invoking recalibration as a major mechanism for accent accommodation.

Acknowledgments

This work was supported by grants from the Economic and Social Research Council (UK) (grant ES/R006288/1) and from the Spanish Ministry for Science and Innovation (grants PSI2017-82563-P and PID2020-113348GB), by the Basque Government, through the BERC 2018-2021 program, and by the Spanish State Research Agency, through the BCBL Severo Ochoa excellence accreditation (grants SEV-2015-0490 and CEX2020-001010-S). In addition, we would like to thank Dr. Susan Brennan, Dr. Marie Huffman and Dr. Toni Freitas for helpful advice on this project.

Open Practice Statement

The data and analyses for all experiments are available at *https://github.com/jeanne-charoy/BadmapProject.*

References

Alderete, J., & Davies, M. (2019). Investigating perceptual biases, data reliability, and data discovery in a methodology for collecting speech errors from audio recordings. *Language and speech*, *62*(2), 281-317.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese/r/and English/l/and/r. *Journal of Phonetics*, *32*(2), 233-250.

Babel, M., Johnson, K. A., & Sen, C., 2021. Asymmetries in perceptual adjustments to non-canonical pronunciations. Laboratory Phonology: Journal of the Association for Laboratory Phonology, 12(1): 19, pp. 1–43. DOI: https://doi.org/10.16995/labphon.6442

Babel, M., McAuliffe, M., Norton, C., Senior, B., & Vaughn, C. (2019). The Goldilocks zone of perceptual learning. *Phonetica*, *76*(2-3), 179-200.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of phonetics*, *20*(3), 305-330.

Best, C. T., Shaw, J. A., Mulak, K. E., Docherty, G., Evans, B. G., Foulkes, P., ... & Wood, S. (2015, August). Perceiving and adapting to regional accent differences among vowel subsystems. In *ICPhS*.

Bion, R. A. H., Escudero, P., & Morrison, G. S. (2008, June). Dialectal effects in the perception of vowels produced by first and second language speakers: North Carolinian versus Southern Welsh listeners. In *Proceedings of Meetings on Acoustics 155ASA* (Vol. 4, No. 1, p. 060005). Acoustical Society of America.

Blanco-Elorrieta, E., Gwilliams, L., Marantz, A., & Pylkkänen, L. (2021). Adaptation to mis-pronounced speech: evidence for a prefrontal-cortex repair mechanism. *Scientific reports*, *11*(1), 1-11.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0. 37. *Retrieved February*, *3*, 2018.

Bohn, O. S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition*, *14*(2), 131-158.

Bradlow, A. R., & Bent, T. (2003). Listener adaptation to foreign-accented English. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2881-2884). Universitat Autònoma de Barcelona.

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

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PloS one*, *13*(8), e0199358.

Burgos, P., Cucchiarini, C., van Hout, R., & Strik, H. (2014). Phonology acquisition in Spanish learners of Dutch: Error patterns in pronunciation. *Language Sciences*, *41*, 129-142.

Carnegie Mellon University Pronouncing Dictionary. (2014). Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Cebrian, J. (2007). Old sounds in new contrasts: L2 production of the English tense-lax vowel distinction. *Poster presented at the 16 th International Congress of Phonetic Sciences, Universitat des Saarlandes, Germany. Retrieved December* (Vol. 30, p. 2007).

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias?. *Perception & psychophysics*, *70*(4), 604-618.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language*, *30*(2), 234-250.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition?. *Journal of Memory and Language*, *32*(2), 193-210.

Connine, C. M., Blasko, D. G., & Wang, J. (1994). Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context. *Perception & Psychophysics*, *56*(6), 624-636.

Cooper, A., & Bradlow, A. R. (2016). Linguistically guided adaptation to foreign-accented speech. *The Journal of the Acoustical Society of America*, *140*(5), EL378-EL384.

Cooper, A., & Bradlow, A. (2018). Training-induced pattern-specific phonetic adjustments by first and second language listeners. *Journal of phonetics*, *68*, 32-49.

Cooper, A., Paquette-Smith, M., Bordignon, C., & Johnson, E. K. (2023). The influence of

accent distance on perceptual adaptation in toddlers and adults. *Language Learning and Development*, *19*(1), 74-94.

Cutler, A., Smits, R., & Cooper, N. (2005). Vowel perception: Effects of non-native language vs. non-native dialect. *Speech communication*, *47*(1-2), 32-42.

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, *19*(2), 121-126.

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

Evanini, K., & Huang, B. (2012, June). Automatic detection of [th] pronunciation errors for Chinese learners of English. In *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training, Stockholm* (pp. 71-74).

Fabra, L. R., & Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, *40*(3), 491-508.

Flege, J. E. (1987). The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of phonetics*, *15*(1), 47-65.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233-277.

Flege, J. E., Munro, M. J., & Skelton, L. (1992). Production of the word-final English/t/–/d/ contrast by native speakers of English, Mandarin, and Spanish. *The Journal of the Acoustical Society of America*, *92*(1), 128-143.

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English-language experience affect Japanese adults’ perception of/ɹ/and/l. *The Journal of the Acoustical Society of America*, *99*(2), 1161-1173.

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.

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, *30*(2), 139-162.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds" L" and" R.". *Neuropsychologia*.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585-7599.

Gwilliams, L., Poeppel, D., Marantz, A., & Linzen, T. (2017). Phonological (un) certainty weights lexical activation. *arXiv preprint arXiv:1711.06729*.

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., & Weber, A. (2010). Production of English interdental fricatives by Dutch, German, and English speakers. In *New Sounds 2010: Sixth International Symposium on the Acquisition of Second Language Speech* (pp. 173-178). Adam Mickiewicz University.

Idemaru, K., Wei, P., & Gubbins, L. (2019). Acoustic sources of accent in second language Japanese Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics*, *37*(2), 303.speech. *Language and speech*, *62*(2), 333-357.

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

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, *119*(2), 1118-1130.

Kleinschmidt, D. F., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34).

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

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

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

Levy, E. S., & Law, F. F. (2010). Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship. *The Journal of the Acoustical Society of America*, *128*(3), 1290-1305.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55-70.

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1562.

Lombardi, L. (2003). Second language data and constraints on manner: Explaining substitutions for the English interdentals. *Second Language Research*, *19*(3), 225-250.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, *10*(1), 29-63.

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.

McAuliffe, M., & Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *The Journal of the Acoustical Society of America*, *140*(3), 1727-1738.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, *60*(1), 65-91.

Mitterer, H. (n.d.) Praat script retrieved from http://www.holgermitterer.eu/research.html

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

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

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

Picard, M. (2002). The differential substitution of English/θ ð/in French: The case against underspecification in L2 phonology. *Lingvisticæ Investigationes*, *25*(1), 87-96.

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.

Rau, D. V., Chang, H. H. A., & Tarone, E. E. (2009). Think or sink: Chinese learners' acquisition of the English voiceless interdental fricative. *Language Learning*, *59*(3), 581-621.

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.

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

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, *110*(4), 474.

Samuel, A. G. (1991). A further examination of attentional effects in the phonemic restoration illusion. *The Quarterly Journal of Experimental Psychology Section A*, *43*(3), 679-699.

Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive psychology*, *88*, 88-114.

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

Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics*, *37*(2), 303.

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

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, *75*(7), 1533-1546.

Teasdale, A. (1997). On the differential substitution of English [θ] A phonetic approach. Calgary Working Papers in Linguistics, 19(Winter), 71-92.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics*, *33*(3), 263-290.

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

Witteman, M. J., Bardhan, N. P., Weber, A., & McQueen, J. M. (2015). Automaticity and stability of adaptation to a foreign-accented speaker. *Language and speech*, *58*(2), 168-189.

Witteman, M. J., Weber, A., & McQueen, J. M. (2010). Rapid and long-lasting adaptation to foreign-accented speech. *The Journal of the Acoustical Society of America*, *128*(4), 2486-2486.

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.

Witteman, M. J., Weber, A., & McQueen, J. M. (2014). Tolerance for inconsistency in foreign-accented speech. *Psychonomic bulletin & review*, *21*(2), 512-519.

Zhang, Y., & Xiao, J. (2014). An Analysis of Chinese Students' Perception and Production of Paired English Fricatives: From an ELF Perspective. *Journal of Pan-Pacific Association of Applied Linguistics*, *18*(1), 171-192.

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

Zheng, Y., & Samuel, A.G. (2023). Flexibility and stability of speech sounds: The time course of lexically-driven recalibration. *Journal of Phonetics*.

Appendix A

Noun task critical items

Early /s/ words	Late /s/ words
Acerbic (əˈsɛrbɪk)	Address (əˈdrɛs)
Basement (ˈbeɪsmənt)	Announcement (əˈnaʊnsmənt)
Casserole (ˈkæsəˌroʊl)	Arkansas (ˈɑrkənˌsɑ)
Cider (ˈsaɪdər)	Arthritis (ɑrˈθraɪtəs)
Gasoline (ˈgæsəˌlin)	Coliseum (ˌkɑləˈsiəm)
Sabre (ˈseɪbər)	Dinosaur (ˈdaɪməˌsɔr)
Salty (ˈsɔlti)	Embarrassing
Sandal (ˈsændəl)	Embassy (ˈɛmbəsi)
Seaweed (ˈsiˌwid)	Embrace (ɛmˈbreɪs)
Secluded (sɪˈkludɪd)	Eraser (ɪˈreɪsər)
Second (ˈsɛkənd)	Hallucinate (həˈlusəneɪt)
Several (ˈsɛvərəl)	Homogeneous
Sidewalk (ˈsaɪˌdwɔk)	Malpractice (mælˈpræktəs)
Silver (ˈsɪlvər)	Miraculous (məˈrækjələs)
Similar (ˈsɪmələr)	Nervous (ˈnɜrvəs)
Solid (ˈsɑləd)	Peninsula (pəˈnɪnsələ)
Submarine (ˈsʌbməˌrin)	Pregnancy (ˈprɛgnənsi)
Suffer (ˈsʌfər)	Rehearsal (rɪˈhɜrsəl)
Superpower (ˌsupərˈpaʊər)	Tennessee (ˌtɛnəˈsi)
Surrender (səˈrɛndər)	Utensil (juˈtensəl)

Appendix B

Questionnaire online

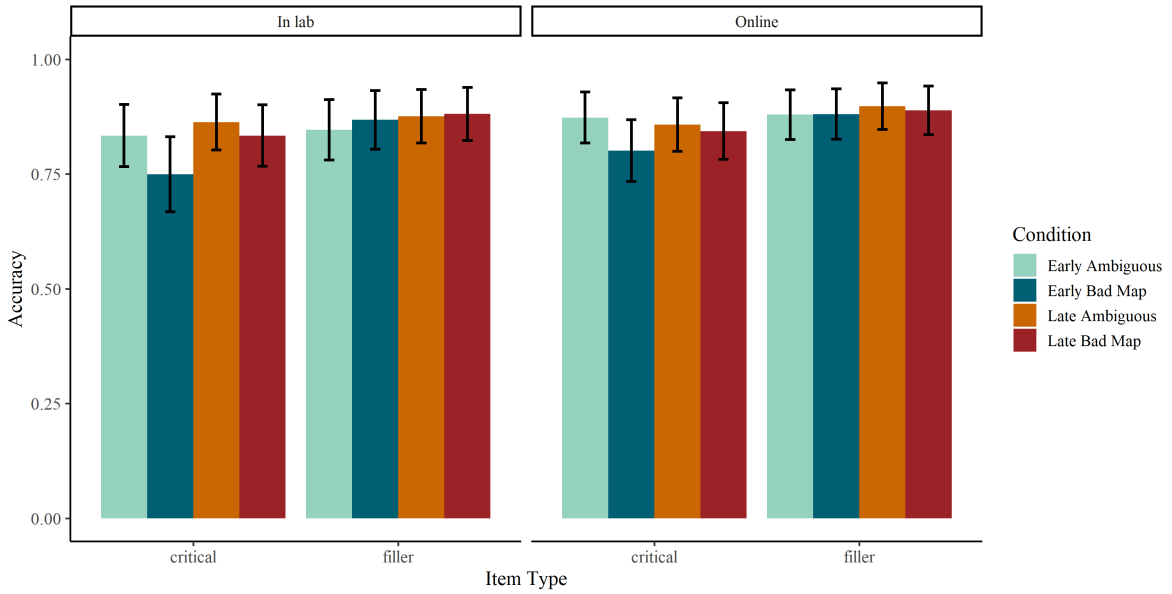
1. Age
2. Sex (Male, Female, Other)
3. Is your hearing normal or corrected to normal (e.g., with hearing aids)?
4. What is the first language you learned as a young child?
    a. English
    b. English and another language at the same time, both from birth
    c. Korean
    d. Mandarin
    e. Spanish
    f. Other – If Other specify
5. At what age did you begin to learn English?
    a. At birth (0-1 year old)
    b. Young childhood (2-5 years old)
    c. Middle childhood (6-12 years old)
    d. Teenage years/adulthood (13 years old or older))
6. What kind of speakers/headphones did you use? Choose the best match to your actual equipment.
    a. Laptop/pc speakers
    b. External speakers (cost $30 or less)
    c. External speakers (cost between $30 and $100)
    d. External speakers (cost more than $100)
    e. In-ear headphones/earbuds (cost $30 or less)
    f. In-ear headphones/earbuds (cost between $30 and $100)
    g. In-ear headphones/earbuds (cost between $30 and $100)
    h. In-ear headphones/earbuds (cost more than $100)
    i. Over-the-ear headphones (cost $30 or less)

j. Over-the-ear headphones (cost between $30 and $100)

k. Over-the-ear headphones (cost more than $100)s

Appendix C

Accuracy results on the Noun task.



Accuracy results for the Noun task for all four conditions across experimental settings (in lab results on the left and online results on the right). There were no significant differences based on condition, suggesting all participants performed similarly on this task and bad map sounds did not hamper performance more than the ambiguous sounds (although note that accuracy was numerically smaller for the Early Bad map condition compared to the other ones, in both versions).

Appendix D

Stimuli selection for the phoneme identification and Noun tasks

Selection of the critical words steps for the Noun task.

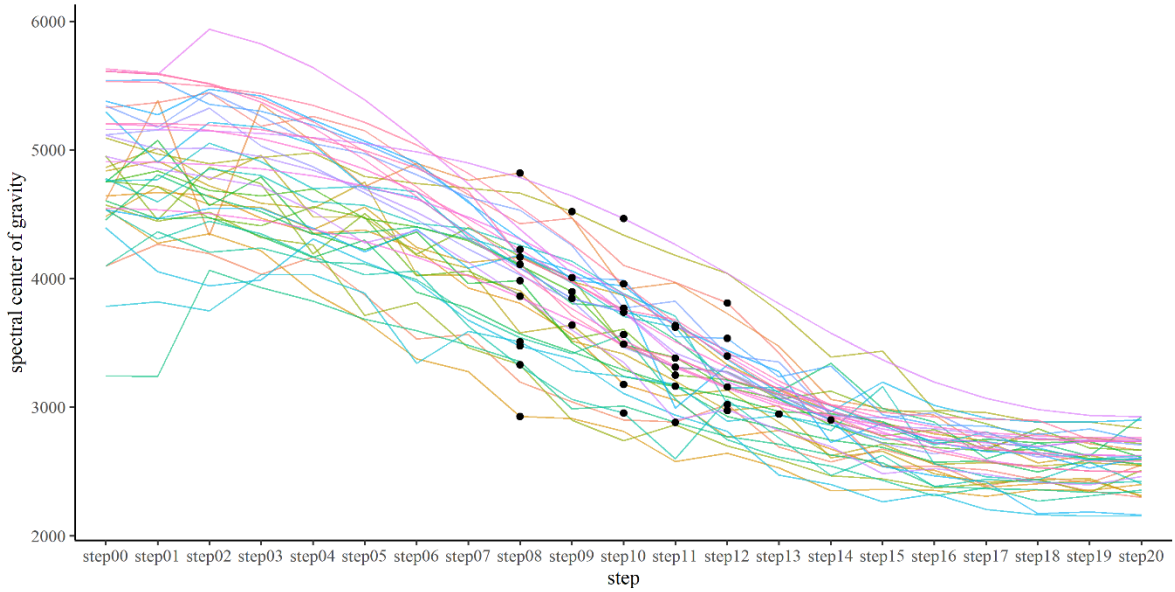
We selected 40 English words that contained the phoneme /s/. In 20 of these words, /s/ occurred early (i.e., before the word's uniqueness point) and in the other 20 it occurred late (i.e., after the uniqueness point). These words were recorded by a male American English speaker both in their correct /s/ form and in an incorrect form where /s/ was replaced with /ʃ/. For each word, the two forms were merged using a PRAAT script (Mitterer, n.d.), resulting in 21-step continua where the fricative were gradient composites of /s/ and /ʃ/ (more details in the main text).

We asked 8 native American English speakers to listen to the stimuli and to select the items they found hardest to categorize as an /s/ word or an /ʃ/ word. So, for example, listeners would hear the 21 versions of the item "superpower" ranging from step 1 (where the /s/ is canonical) to step 21 (where it is fully replaced with an /ʃ/) and select which middle step they found most ambiguous. This is a typical procedure for stimulus selection in perceptual recalibration experiments and it has been used previously in our lab (e.g., Kraljic & Samuel, 2005). The 40 words were divided into two lists and each list was rated by a subset of 4 of the native American English speakers (so each word was rated by 4 people). This was done to limit the amount of work asked of each rater. The complete list of ratings (with each rater's judgments) can be found at *https://github.com/jeanne-charoy/BadmapProject*. The average ratings for each word are summarized in a table below.

Word	Most Ambiguous Step	Word	Most Ambiguous Step
Acerbic	12	Peninsula	11
Address	12	Pregnancy	8
Announcement	8	Rehearsal	8
Arkansas	10	Sabre	10
Arthritis	8	Salty	14
Basement	11	Sandal	11
Casserole	12	Seaweed	12
Cider	9	Secluded	10
Coliseum	9	Second	11
Dinosaur	8	Several	9
Embarrassing	9	Sidewalk	8
Embassy	11	Silver	10
Embrace	9	Similar	10
Eraser	8	Solid	12
Gasoline	11	Submarine	10
Hallucinate	8	Suffer	8
Homogeneous	11	Superpower	12
Malpractice	10	Surrender	8
Miraculous	10	Tennessee	11
Nervous	13	Utensil	10

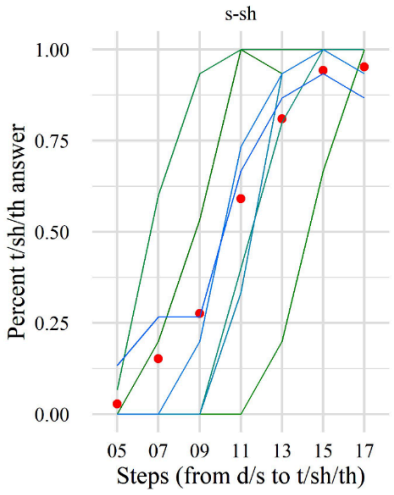
In addition, we used a PRAAT script created by DiCanio (2013) to measure the spectral moments of the fricatives in every step of each of the critical words (so 21 steps times 40 words). This allowed us to match the native speakers' perceptual judgments with acoustic values (specifically the spectral center of gravity of the fricatives, which is one important differentiator between /s/ and /ʃ/ sounds). Overall, listeners' judgments match with the "acoustical middle". A

graph representing these two measures (perceptual judgments and spectral center of gravity values) is found below.



Selection of the ASEE - ASHEE steps for the phoneme categorization task.

The continua used to evaluate phoneme categorization were normed based on the responses of five native American English speakers. The norming aimed to ensure that the step chosen would span the continuum, including some clear tokens (e.g., a clear “s”) and some highly ambiguous ones. On the figure above, each line corresponds to the average responses of one rater, while the red dots represent the group’s mean.



Appendix E.

Full analyses for the Phoneme Categorization task

We used generalized mixed-effect models to compare performance in the phoneme categorization task across the four experimental conditions (Late Ambiguous, Early Ambiguous, Late Bad Map and Early Bad Map) and the Baseline, with separate analyses run for the in-lab and for the online versions. Models included the fixed factors Condition (Baseline was the reference level), Step (centered on 0) and Trial. They also included random intercepts for Participant and random slopes for Step over Participant. Note that only results on the five middle steps of the /s/-/ʃ/ continuum were included in the analysis. A significant difference between an experimental condition and the baseline, with higher percentage of S answers in the former, would indicate the expected boundary shift. First, we present results for each condition when taking the full dataset into account. Next, we present results when considering only the last five repetitions of the continuum steps. Results for the first five repetitions are presented in the main text.

When the data from all of the trials were considered, we observed a trend towards the expected boundary shift for the Late Ambiguous condition in the lab ($b = 0.443$, $SE = 0.233$, *Wald's z* = 1.90, $p = 0.058$) and no effects in the other conditions (Early Ambiguous, $p = 0.23$; Late Bad Map, $p = 0.613$; Early Bad Map, $p = 0.163$). In contrast, all conditions produced a significant boundary shift for online participants: Late Ambiguous ($b = 0.48$, $SE = 0.23$, *Wald's z* = 2.09, $p = 0.036$), Early Ambiguous ($b = 0.56$, $SE = 0.23$, *Wald's z* = 2.41, $p = 0.016$), Late Bad

Map (b = 0.66, SE = 0.23, *Wald's* z = 2.87, p = 0.004) and Early Bad Map (b = 0.49, SE = 0.23, *Wald's* z = 2.13, p = 0.033).

Results that focused on the first five repetitions of each step of the /s/-/ʃ/ continuum are presented in the main text. In short, we found that Late Ambiguous, Early Ambiguous and Early Bad Map conditions led to significant or trending effects in the expected direction in the lab. Online, all conditions produced a significant effect. For the last five repetitions of the phoneme categorization task, recent findings (Liu & Jaeger, 2018, 2019) suggest that recalibration effects may fade towards the end of the task. Indeed, this is what we find both in lab and online[1] for all conditions: Late Ambiguous (in lab, p = 0.90; online, p = 0.72); Early Ambiguous (in lab, p = 0.57; online, p = 0.38); Late Bad Map (in lab, p = 0.38; online, p = 0.084) and Early Bad Map (in lab, p = 0.45; online, p = 0.41). These results clearly support the idea that exposure to the continuum steps during the phoneme categorization task itself eventually wipes out the recalibration effect.

We conducted a similar analysis for our entire dataset (i.e., including both experimental settings) focusing on the first five repetitions of the task. This provides us with unusually large sample sizes for such experiments (i.e., 63 to 68 participants for the experimental conditions and 144 for the Baseline). When pooling the data in this way, we find significant boundary shifts for all conditions, with the largest effect for the Late Ambiguous condition (b = 0.88, SE = 0.192, *Wald's* z = 4.58, $p < 0.001$), followed by the Early Bad Map condition (b = 0.638, SE = 0.197, *Wald's* z = 3.25, p = 0.001), the Early Ambiguous condition (b = 0.62, SE = 0.194, *Wald's* z =

[1] Note that for the online data analyses here the random slope for Step over Participant was not included because it prevented the model from converging.

3.20, $p = 0.001$) and the Late Bad Map condition ($b = 0.53$, $SE = 0.193$, *Wald's* $z = 2.73$, $p = 0.0062$). Means are summarized in the table below.

Condition	% S answer	n
Baseline	34.9%	144
Late Ambiguous	45.2%	68
Early Ambiguous	42.3%	66
Late Bad Map	41.1%	66
Early Bad Map	42.7%	63

Appendix F

Full descriptions of the statistical models.

For all of our analyses we used generalized linear mixed model fit by maximum likelihood (Laplace Approximation), with a binomial family. This was to account for our binomial dependent variables (Grammatical Type in the Noun task, with values ‘noun’ and ‘not noun’; Phoneme identity in the Phoneme identification task, with values ‘S’ or ‘SH’). Here we list the models’ output for each analyses presented in the text, starting with the in-lab dataset, then the online one and finally the cross-experiment results.

In-Lab experiment analysis.

Noun task.

Model: Accuracy ~ Condition + Item Type + Trial + (1|Item) + (1 + Item Type + Grammar Type | Participant)

Random effects:

Group	Name	Variance	Standard Deviation	Correlation	
Item	(Intercept)	2.605	1.614		
Participants	(Intercept)	0.972	0.986		
	Item type (filler vs. critical)	0.177	0.421	-0.36	
	Grammar type (noun vs non noun)	3.793	1.948	-0.87	0.36

Fixed effects (reference level is the Late Ambiguous condition):

	Estimate	Std. Error	Z value	Pr(>|z|)

(Intercept)	2.6222	0.3081	8.51	< 0.001
Early Ambiguous	- 0.3097	0.1852	-1.67	0.095
Late Bad Map	- 0.0362	0.1841	0.20	0.84
Early Bad Map	- 0.0060	0.1932	- 0.03	0.975
Item Type (Filler)	0.4332	0.2798	1.55	0.122
Grammar Type (Noun)	- 0.1629	0.3316	- 0.49	0.623
Trial	- 0.0419	0.0322	- 1.30	0.193

Phoneme Categorization task:

Model: Percent S Answer ~ Condition + Step+ Trial + (1 + Step | Participant)

Random effects:

Group	Name	Variance	Standard Deviation	Correlation
Participants	(Intercept)	1.18	1.08	
	Step	6.23	2.50	0.19

Fixed effects (reference level is the Baseline condition):

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	- 1.2412	0.1611	-7.70	< 0.001 ***
Late Ambiguous	0.9704	0.2689	3.61	< 0.001 ***
Early Ambiguous	0.4997	0.2759	1.81	0.07012 .

Late Bad Map	0.2203	0.2683	0.82	0.4116
Early Bad Map	0.6312	0.2831	2.23	0.02576 *
Step	- 6.8137	0.2922	-23.32	<0.001 ***
Trial	0.0174	0.0446	0.39	0.6966

Phoneme Categorization task - Interaction between Position and Pronunciation :

Model: Percent S Answer ~Pronunciation x Position + Step + Trial + (1 + Step | Participant)

Group	Name	Variance	Standard Deviation	Correlation
Participants	(Intercept)	11.5056	3.392	
	Step	0.0973	0.312	-0.95

Fixed effects (reference level is the Baseline condition):

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	- 1.2412	0.1611	-7.70	< 0.001
Pronunciation (ambiguous)	0.9704	0.2689	3.61	< 0.001
Pronunciation (bad map)	0.2202	0.2683	0.82	0.41163
Position (early)	0.4410	0.3249	1.27	0.20586
Step	-6.8137	0.2922	-23.32	<0.001
Trial	- 6.8137	0.0446	0.39	0.6966
Pronunciation x Position	- 0.8817	0.4529	-1.95	0.05154

Online experiment analysis.

Noun task.

Model: Accuracy ~ Condition + Item Type + Grammar Type + Trial + (1|Item) + (1 + Item Type + Grammar Type | Participant)

Random effects:

Group	Name	Variance	Standard Deviation	Correlation	
Item	(Intercept)	2.756	1.660		
Participants	(Intercept)	0.949	0.974		
	Item type (filler vs. critical)	0.182	0.427	-0.18	0.42
	Grammar Type	4.012	2.003	-0.92	

Fixed effects (reference level is the Late Ambiguous condition):

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	3.2251	0.3193	10.10	< 0.001
Early Ambiguous	-0.0891	0.2013	-0.44	0.658
Late Bad Map	-0.2271	0.1990	-1.14	0.254
Early Bad Map	-0.2476	0.2004	-1.24	0.217
Item Type (Filler)	0.5329	0.2869	1.86	0.063 .
Grammar Type (Noun)	-0.6956	0.3350	-2.08	0.038 *
Trial	-0.0233	0.0327	-0.71	0.475

Phoneme Categorization task:

Model: Percent S Answer ~ Condition + Step+ Trial + (1 + Step | Participants)

Random effects:

Group	Name	Variance	Standard Deviation	Correlation
Participants	(Intercept)	11.115	3.334	
	Step	0.0774	0.278	-0.93

Fixed effects (reference level is the Baseline condition):

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-1.4343	0.1653	-8.68	< 0.001
Late Ambiguous	0.8561	0.2766	3.09	0.002
Early Ambiguous	0.8252	0.2808	2.84	0.0033
Late Bad Map	0.8299	0.2794	2.97	0.003
Early Bad Map	0.7716	0.2790	2.77	0.0057
Step	-6.6734	0.2592	-25.75	<0.001
Trial	0.5606	0.0442	12.68	<0.001

<u>Cross Experiment analysis.</u>

<u>Noun Task.</u>

Model: Accuracy ~ Condition + Item Type + Grammar Type + Trial + (1|Item) + (1 + Item Type + Grammar Type | Participant):

Random effects:

Group	Name	Variance	Standard Deviation	Correlation	

Participants	(Intercept)	0.949	0.974		
	Item type (filler vs. critical)	0.182	0.427	-0.18	0.42
	Grammar Type	4.012	2.003	-0.92	
Item	(Intercept)	2.045	1.430		

Fixed Effects (reference level was the Late Ambiguous Condition).

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	2.6097	0.2511	10.39	<0.001
Early Ambiguous	-0.1804	0.1368	-1.32	0.187
Late Bad Map	-0.0361	0.1372	-0.26	0.793
Early Bad Map	-0.0806	0.1390	-0.58	0.562
Experiment (1B)	0.1651	0.0911	1.81	0.07
Item Type (Filler)	0.3376	0.2391	1.41	0.158
Grammar Type (Noun)	-0.5092	0.2370	-2.15	0.032
Trial	-0.0356	0.0216	-1.64	0.10

Perceptual Recalibration Task.

Model: Percent S Answer ~ Condition x Experiment + Step+ Trial + (1 | Participants)

Note that a random slope for Step on Participants was not included because it prevented the model from converging.

Random effects :

Group	Name	Variance	Standard Deviation	Correlation	
Participants	(Intercept)	1.22	1.11		

Fixed effects :

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-0.7702	0.1549	-4.97	<0.001
Late Ambiguous	0.9513	0.2681	3.55	<0.001
Early Ambiguous	0.433	0.2738	1.58	0.1338
Late Bad Map	0.2082	0.2708	0.77	0.442
Early Bad Map	0.5377	0.2804	1.92	0.0551
Experiment (Online)	-0.9217	0.2168	-4.25	<0.001
Step	-6.0225	0.1212	-49.67	<0.001
Trial	0.5	0.0395	12.64	<0.001
Late Ambiguous x Experiment	-0.1337	0.3687	-0.36	0.7168
Early Ambiguous x Experiment	0.3704	0.3735	0.99	0.3214
Late Bad Map x Experiment	0.6113	0.3725	1.64	0.1007
Early Bad Map x Experiment	0.2209	0.3794	0.58	0.5603