Published in final edited form as:

Cognition. 2018 May ; 174: 55–70. doi:10.1016/j.cognition.2018.01.003.

Inferring causes during speech perception

Linda Liu[1] and **T. Florian Jaeger**[1,2]

[1]Department of Brain and Cognitive Sciences, University of Rochester, USA

[2]Department of Computer Science, University of Rochester, USA

Abstract

One of the central challenges in speech perception is the lack of invariance: talkers differ in how they map words onto the speech signal. Previous work has shown that one mechanism by which listeners overcome this variability is adaptation. However, talkers differ in how they pronounce words for a number of reasons, ranging from more permanent, characteristic factors such as foreign accent, to more temporary, incidental factors, such as speaking with a pen in the mouth. One challenge for listeners is that the true cause underlying atypical pronunciations is never directly known, and instead must be inferred from (often causally ambiguous) evidence. In three experiments, we investigate whether these inferences underlie speech perception, and how the speech perception system deals with uncertainty about competing causes for atypical pronunciations. We find that adaptation to atypical pronunciations is affected by whether the atypical pronunciations are seen as characteristic or incidental. Furthermore, we find that listeners are able to maintain information about previous causally ambiguous pronunciations that they experience, and use this previously experienced evidence to drive their adaptation after additional evidence has disambiguated the cause. Our findings revise previous proposals that causally ambiguous evidence is ignored during speech adaptation.

Keywords

speech perception; perceptual recalibration; talker variation; accent adaptation; causal reasoning

1 Introduction

One of the fundamental challenges in speech perception is the lack of invariance in the mapping from acoustic inputs to phonological categories (e.g., the phoneme /s/ as in *sip*). Due to factors ranging from anatomical differences (e.g., vocal tract length or vocal fold size) to social factors (e.g., dialects or accents), the mapping from acoustic signal to phonological category varies from talker to talker (see Klatt, 1986 for an overview). Understanding how the systems underlying human speech perception typically overcome

Corresponding Author: Linda Liu, Department of Brain and Cognitive Sciences, University of Rochester, lliu47@ur.rochester.edu.

such talker-specific differences continues to be one of the central problems in research on speech perception (for review, see Pardo & Remez, 2006; Weatherholtz & Jaeger, 2016).

This research has highlighted adaptation—the ability to adjust phonological categories through some form of implicit learning—as one important mechanism that allows listeners to overcome inter-talker variability (e.g. Bradlow & Bent, 2008; Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2003; Sidaras, Alexander, & Nygaard, 2009, for a review of this and other mechanisms, see Weatherholtz & Jaeger, 2016). For example, when listeners are first exposed to a novel dialect or foreign accent, they may initially experience processing difficulty, as reflected in slower processing speeds and lower comprehension accuracy (Bradlow & Bent, 2008; Clarke & Garrett, 2004). However, this difficulty can rapidly decrease with additional exposure (ibid; see also Baese-Berk, Bradlow, & Wright, 2013; Nygaard, Sommers, & Pisoni, 1994; Sidaras et al., 2009). Similarly rapid adaptation to *a priori* unexpected pronunciations has also been observed for less pervasive talker-specific differences, such as shifted pronunciations of individual phonemes (e.g. Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Norris et al., 2003; Vroomen, van Linden, De Gelder, & Bertelson, 2007).

Adaptation thus forms an important part of how listeners overcome inter-talker-variability: when successful, adaptation to talker-specific pronunciations facilitates perception of future productions by the same talker. However, this seemingly innocuous statement hides an important complexity that has so far received relatively little attention: successful adaptation requires listeners to distinguish between pronunciations that are *characteristic of the talker*—i.e., informative about their future productions—and those that are not.[1] How listeners accomplish this is the question we seek to contribute to here.

Not all atypical pronunciations are characteristic of the *talker*, or, more specifically, not all atypical pronunciations are equally informative about future pronunciations by the same talker. On a first encounter, a talker might, for example, be in a particular physical or emotional state known to temporarily influence pronunciations, such as being under the influence of alcohol (Chin & Pisoni, 1997; Johnson, Pisoni, & Bernacki, 1990; Pisoni & Martin, 1989) or in a strong emotional state (Sobin & Alpert, 1999; Williams & Stevens, 1972). Atypical pronunciations can even arise from entirely incidental causes, such as talking with the mouth full or having a pen in the mouth. Any of these states are less informative about the talker than more permanent properties, such as the talker's dialect background or vocal tract length. In short, a variety of causes affect our speech, and these causes differ in how predictive they are about future pronunciations by the same talker.

Consequently, it would theoretically be advantageous for listeners to take the cause for an observation into consideration when adapting to a talker. The more likely the cause of an observed pronunciation is to be present on future encounters with the same talker—i.e., the more characteristic the pronunciation is for the talker—the more it will help to adapt to and store this pronunciation information as part of talker-specific knowledge. This would allow

[1] Alternatively, we might describe this problem as determining which memories or representations based on previous experiences are relevant to the interpretation of the current input (see Kleinschmidt & Jaeger, 2015). We return to this point in the discussion.

listeners to generalize their previous experience more effectively to future encounters with the same talker (or similar talkers). Adapting indiscriminately, on the other hand, would risk unnecessarily volatile speech perception (Kleinschmidt & Jaeger, 2015; Samuel, 2011). For example, drunkenness may result in changes to how a talker produces /s/ sounds, and a listener who fails to recognize this as a situation-specific shift may have trouble categorizing this sound correctly on future encounters with the same (but sober) talker (Kleinschmidt & Jaeger, 2015; for the more general problem of learning under non-stationary statistics, see also Qian, Jaeger, & Aslin, 2012; Yu & Cohen, 2008).

However, causes of atypical pronunciations are not directly observable to listeners. It is only the evidence for a cause that is observable (e.g., a pen in the talker's mouth or visible signs that the talker is drunk). The cause itself needs to be *inferred.* Often there will be several possible causes, thus rendering the perceptual evidence *causally ambiguous.* For example, imagine that you encounter a novel talker who produces an atypical /s/ while chewing on a pen. While the atypical /s/ could be due to having a pen in the mouth (Kraljic, Samuel, & Brennan, 2008), it could alternatively be due to the talker's lisp (making it characteristic of the talker) and be largely unrelated to the pen's presence. In everyday speech perception, there will be uncertainty about the true cause(s) underlying an atypical pronunciation.

Here, we ask whether listeners indeed draw on inferences about the causes of unexpected pronunciations (henceforth, causal inferences), and if so, how they deal with causal uncertainty about the atypical pronunciations. On the one hand, listeners may draw inferences about the cause(s) for a talker's atypical pronunciations, affecting how they generalize these pronunciations to future input from the same talker. On the other hand, this type of inference may be too complex under the demands inherent to speech perception, leaving listeners to indiscriminately adapt to any input they receive. Indeed, most existing models of implicit learning during language processing assume that comprehenders indiscriminately adapt to any input they receive (e.g., Chang, Dell, & Bock, 2006; Dell & Chang, 2014; Lancia & Winter, 2013; Reitter, Keller, & Moore, 2011; but see Kleinschmidt & Jaeger, 2015), although we stress that these models were not developed to address the question raised here.

If listeners indeed engage in causal inference, this raises questions about the nature of the mechanisms that support these inferences. Given causally ambiguous evidence of an atypical pronunciation, do listeners maintain uncertainty about the potential causes (i.e. pen in mouth or lisp)? If so, if later provided with disambiguating evidence (i.e. strong evidence that the true cause is a pen in the mouth), can listeners integrate this evidence with previously observed causally ambiguous pronunciations, thereby disambiguating the cause(s) of these previous pronunciations and functionally rendering them causally unambiguous? This is beneficial because causally unambiguous pronunciations provide unambiguous evidence to adapt or to not adapt. Such a system would, however, also place higher memory demands on speech perception, as it would require listeners to store subcategorical information about percepts along with the context the percepts were experienced in.

We begin our study with a summary of what is known about the role of alternative causes for atypical pronunciations during speech adaptation. Despite the potentially central importance of causal inferences to speech perception (see above), only two studies—discussed below—have addressed these questions (Kraljic & Samuel, 2011; Kraljic et al., 2008). We first discuss these studies in some detail, as they provide the motivation for the present work. The experimental paradigm we employ below closely builds on these studies. We then discuss why previous evidence taken to argue *against* the hypothesis that speech adaptation draws on causal inferences (Kraljic & Samuel, 2011) is in fact compatible with this hypothesis. This leads us to conduct a series of experiments in which we investigate speech perception and adaptation under uncertainty about the cause(s) of atypical pronunciations.

1.1 Previous Work and its Interpretation

Kraljic and colleagues (2008, 2011) employ a *perceptual recalibration* paradigm to study how alternative causes for atypical pronunciations affect subsequent perception of talker's speech. Perceptual recalibration experiments measure how listeners change their interpretation of a talker's speech following exposure to that talker's atypical pronunciation of a particular phonemic contrast (here, /s/ and /ʃ/). As illustrated in Figure 1, the paradigm consists of an exposure and a test block. During exposure, participants hear individual words (or nonce words) and perform a lexical decision task. On critical trials, the atypical sound halfway between a canonical /s/ and /ʃ/ sound replaces either the /s/ or /ʃ/ sound (between participants). The atypical sound is labeled as either /s/ (S-Label condition) or /ʃ/ (ʃ-Label condition) based on lexical context. For example, participants might hear the word "dinosaur" pronounced halfway between "dinosaur" and "dinoshaur" in the S-Label condition.

During test, participants categorize nonce words along an /asi/-/aʃi/ continuum as either /asi/ or /aʃi/. This type of study has found that the S-Label group expands their /s/-category towards /ʃ/, now categorizing more sounds as /s/ compared the ʃ-Label group (e.g., /s/ vs. /ʃ/: Kraljic & Samuel, 2005; /f/ vs. /s/: Norris et al., 2003; /r/ vs. /l/: Scharenborg, Mitterer, & McQueen, 2011). This is illustrated on the right hand side of Figure 1.

Kraljic and colleagues (2008, 2011) used this paradigm to study whether perceptual recalibration is sensitive to the presence of a plausible alternative cause. Figure 2 provides a summary of the experiments presented in those two papers, explained in more detail below. We begin with Experiment 1a from Kraljic et al. (2008). This experiment closely resembles the standard perceptual recalibration design with one exception: half of the critical stimuli (either /s/ or /ʃ/) were atypical pronunciations, while the other half were normal pronunciations. Between participants, Kraljic and colleagues manipulated whether the atypical or normal pronunciations occurred in the first or second half of the exposure block. Kraljic and colleagues found adaptation—i.e., shifted category boundaries, depending on whether atypical stimuli were lexically labeled as /s/ or /ʃ/—only when the atypical sounds occurred in the first half of exposure, i.e. only when normal followed atypical, but not vice versa. They referred to this as the "first impression" effect: listeners adapt when they first encounter a novel talker, but later adaptation to changes within the same talker proceeds more slowly or not at all (for further discussion, see Kleinschmidt & Jaeger, 2015). Kraljic and colleagues use this first impression effect to investigate how listeners integrate causally ambiguous evidence.

In another experiment, Kraljic et al (2008, Experiment 1b) exposed listeners to audiovisual input (videos). Critically, during the first half of the exposure block, participants heard atypical /s/ or /ʃ/ pronunciations that were paired with videos where the talker had a pen either in her mouth or in her hand. The pen in the mouth constitutes a plausible incidental cause (making the evidence *causally ambiguous*, in our terminology). The pen in the hand, on the other hand, provides no plausible incidental cause, suggesting that the pronunciation is likely representative of the talker (and, in this sense, *causally unambiguous*). The manipulation of label (/s/ vs. /ʃ/) and cause (incidental vs. not) occurred between participants. In the second half of the exposure block, both /s/ and /ʃ/ pronunciations were normal and always occurred when the speaker had the pen in the hand.

Consistent with the first impression effect observed in Experiment 1a, Kraljic and colleagues found that listeners adapted to the atypical pronunciations in the first half of exposure when the second half of exposure only contained *normal* pronunciations with hand videos. Crucially, this perceptual recalibration only occurred when the atypical pronunciations in the first half were paired with pen-in-hand videos (and not pen-in-mouth videos). However, perceptual recalibration was again observed when listeners heard normal pronunciations paired with mouth videos followed by atypical pronunciations paired with hand videos (Kraljic & Samuel, 2011, Experiment 3). Thus, the first impression effect thus seems to be only observed for causally *un*ambiguous (pen-in-hand) percepts, and not for causally ambiguous percepts (pen-in-mouth), be they normal or atypical.

Kraljic and Samuel (2011) take this to argue that causally ambiguous percepts are ignored. Specifically, they propose that pronunciations paired with an audiovisual incidental cause are stored separately from those pronunciations without such a cause (ibid, p. 464).[2] According to this perspective, listeners thus effectively *ignore* causally ambiguous percepts, at least while categorizing auditory-only percepts during test.

To further test their proposal that causally ambiguous percepts are ignored, Kraljic and Samuel (2011) conducted two additional experiments (Figure 2). Their first experiment replicates the first impression effect for audio-visual stimuli. In their second experiment, two groups of listeners are first exposed to atypical pronunciations paired with pen-in-mouth videos. In the second half of exposure, listeners then heard atypical percepts that were either paired with pen-in-hand videos or as auditory-only stimuli. In both cases, Kraljic and Samuel observed perceptual recalibration.

Kraljic and Samuel (2011, p. 462) take these results to be "inconsistent with the attribution view of the pen's role [causal inference]. Instead, they indicate that a percept without a pen in the speaker's mouth is encoded as a different type of episode than a percept with the pen." Here we offer an alternative explanation. We argue that all previous findings—including the last experiment discussed above—are perfectly compatible with the idea that speech

[2] Kraljic and Samuel (2011) do not explicitly state whether they assume the latter to be stored together with auditory-only percepts. This assumption is, however, critical. Without it, Kraljic and Samuel's proposal would fail to explain how audiovisual percepts would ever affect categorization along *auditory-only* continua during the test block—contrary to the results of Kraljic et al. (2008; see also Kraljic & Samuel, 2011, Experiment 2). We stress this point because one might argue that this type of separate storage for *specific types* of audiovisual percepts itself constitutes a form of causal inference or attribution. We return to this point in the discussion.

perception can involve causal inferences, *if one keeps in mind that these inferences take place under uncertainty about the true cause of the observed events.* Kraljic and colleagues observe perceptual recalibration when causally ambiguous atypical pronunciations are followed with causally *un*ambiguous atypical pronunciation (Kraljic et al., 2008, Experiment 1b), but not when they were followed with causally *un*ambiguous *normal* pronunciations (Kraljic & Samuel, 2011, Experiment 2). This is straightforwardly accounted for if listeners maintain uncertainty about the cause for the atypical pronunciations—and thus about the true talker-specific characteristics—when presented with causally ambiguous percepts (i.e., during the first half of exposure). The causally unambiguous percepts in the second half of the experiment are then taken to be characteristic of the talker. These are the percepts that determine how listeners will categorize speech from the same talker during test—i.e., whether the listener exhibits perceptual recalibration or not. The same line of reasoning explains why perceptual recalibration is observed after causally unambiguous atypical pronunciations, even when these pronunciations follow causally *ambiguous normal* pronunciations (Kraljic & Samuel, 2011, Experiment 3). This, too, is compatible with the alternative perspective we propose: listeners maintain uncertainty about the true talker-specific characteristics when the initial evidence they observe from a newly encountered talker is causally ambiguous.

1.2 Overview of present study

In summary, existing evidence does not distinguish between two competing views of how listeners treat causally ambiguous atypical speech input. First, listeners might *ignore* any causally ambiguous percepts for atypical pronunciations. This could be the case if listeners completely discarded all information from causally ambiguous percepts, or if they encoded them separately from causally unambiguous percepts, such that they are not used during subsequence categorization. Under this explanation, listeners treat causally ambiguous percepts exactly as if they had not seen these percepts at all. Alternatively, listeners might *maintain* information about the causally ambiguous percepts when they are uncertain as to whether to attribute the pronunciation to characteristic or incidental causes.

Crucially, this latter view predicts that causally ambiguous percepts actually inform subsequent categorization, if later evidence disambiguates the interpretation of the causally ambiguous percepts. Once disambiguated, the previously causally ambiguous percepts now became part of the evidence that listeners take into consideration when categorizing subsequent speech input. If, on the other hand, causally ambiguous evidence is simply ignored, no such prediction is made.

We test this prediction in three experiments. In Experiment 1, we show that listeners do appear to make causal attributions when adapting to atypical pronunciations, such that perceptual recalibration is blocked when the atypical pronunciations are causally ambiguous and plausibly could be attributed to an incidental cause. In Experiments 2 and 3, we test whether listeners can maintain uncertainty about the cause of the atypical pronunciation. We do so by comparing recalibration after equal amounts of exposure to atypical pronunciations that are either always causally unambiguous or initially causally ambiguous, but then causally disambiguated.

2 Experiments 1a and 1b

We first ask whether perceptual recalibration is blocked in the presence of an *a priori* plausible incidental cause. Our experiments closely follow Kraljic et al. (2008, described in the introduction), with one important modification. Kraljic et al. (2008) initially exposed participants to either causally ambiguous (pen-in-mouth) or causally unambiguous (pen-in-hand) atypical pronunciations, followed by causally unambiguous normal pronunciations in both cases. We expose participants exclusively to (either causally ambiguous or unambiguous) atypical pronunciations.[3] Unlike in Kraljic et al. (2008), listeners in the present causally ambiguous condition thus *never* receive disambiguating evidence about the cause for the atypical pronunciation. This provides a baseline for our Experiments 2 and 3, where listeners will receive disambiguating evidence. Additionally, although this is not our primary goal, the design of Experiment 1 allows us to assess participants' prior beliefs about how likely a pen in the mouth is to cause an atypical /s/, compared to an atypical /ʃ/. We elaborate on this point below.

Our experiments also contain some minor procedural changes compared to Kraljic et al. (2008). All of the experiments reported here were conducted over the web on Amazon Mechanical Turk, and use a shorter exposure block. Previous experiments suggest that most perceptual recalibration occurs in the first few trials of exposure (Kleinschmidt & Jaeger, 2011; Vroomen et al., 2007), and shorter paradigms reduce the likelihood of participants losing interest or attention. A similar web-based paradigm was previously successfully used to study perceptual recalibration of /b/ vs. /d/ (Kleinschmidt & Jaeger, 2012; Kleinschmidt & Jaeger, 2015).

Given that our paradigm differs from Kraljic et al. (2008) in a few aspects, we first demonstrate it can in fact detect perceptual recalibration on the /s/ and /ʃ/ contrast in response to causally unambiguous percepts (Experiment 1a). This has the additional advantage of revealing an important property of perceptual recalibration data, leading us to propose a novel analysis that we will then use for the remainder of the paper. We then test whether such perceptual recalibration is weakened following exposure to causally ambiguous percepts (Experiment 1b).

2.1 Method

2.1.1 Participants.—We recruited 224 participants for Experiment 1. An additional 60 participants were recruited for a baseline measure, described later. The experiment took up to 15 minutes and participants were paid $1.50 ($6/hour). Participants were self-reported native speakers of English, and were instructed to complete the experiment while wearing headphones in a quiet room.

The targeted number of participants was determined based on *a priori* power considerations, based on the work we aimed to replicate (Kraljic et al., 2008) and similar web-based

[3] We use the terms causally ambiguous and unambiguous as convenient labels for our conditions. We note, however, that the causes for an observed pronunciation are unlikely to ever be truly unambiguous.

perceptual recalibration studies (Kleinschmidt & Jaeger, 2012). Across all experiments, we aimed to hold the number of participants per between-participant condition near-constant.

2.1.2 Materials.

2.1.2.1 Exposure (Audiovisual) – Lexical Decision.: Participants were assigned to one of eight lists. Four lists were created by Latin square design over two design factors: whether the /s/ or /ʃ/ words were atypical or normal pronunciations, and whether the atypical pronunciations were paired with pen-in-hand (causally unambiguous) or pen-in-mouth (causally ambiguous) videos. The same pseudo-randomized stimulus order was used for all four lists. Four additional lists were created by reversing the stimulus order.

Each list contained 50 words and 50 nonce words, all 2-4 syllables long (Appendix A). Of these, 10 words contained an /s/ sound and 10 contained an /ʃ/ sound. For half of the lists, the /s/ words were shifted towards /ʃ/, such that they contained an ambiguous sound (/~sʃ/) (S-Label condition). For the other lists, the /ʃ/ words were shifted towards /s/ (ʃ-Label condition). No filler words or non-words contained the /s/ or /ʃ/ sound. The words constituted a proper subset (50%) of those used in Kraljic et al. (2008).

The audio files used during exposure were identical to the ones used in Kraljic et al. (2008). All stimuli were produced by the same young, female speaker, including the two endpoints that were blended to create the atypical pronunciations (i.e. *episode* and *epishode*). For additional details about the creation of the atypical pronunciations used during exposure, see Kraljic and Samuel (2005).

Since the original video files of Kraljic et al. (2008) are no longer available (Arthur Samuel, p.c., 11/3/2014), we used the newly recorded videos originally developed for Babel (2016), generously provided by Molly Babel. In each video, the same young, female speaker was seen sitting in front of a green background producing a word while fiddling with a pen, either held in her hand or mouth (see Figure 3).

In Experiment 1a, each of the 10 words containing /~sʃ/ were paired with videos where the speaker held and fiddled with a pen in her *hand* (causally unambiguous; henceforth 10CU). In Experiment 1b, they were paired with videos where the speaker held and fiddled with a pen in her *mouth* (causally ambiguous; henceforth 10CA). The normally pronounced critical tokens were always paired with the pen-in-hand videos. Crucially, the same audio file was used to dub each pair of video files for every word, such that participants in both experiments heard the same audio percepts. For the baseline condition, participants only heard normal pronunciations of the critical stimuli, paired with the pen in the hand videos.

2.1.2.2 Test (Audio only) – Category identification.: The same speaker who produced the exposure stimuli also produced the test stimuli. For the test stimuli, we created a 31-step continuum from /aʃi/ to /asi/ using FricativeMakerPro (McMurray, Rhone, & Galle, 2012). Through a series of three pilot experiments, reported in Appendix C, we identified six steps in this continuum to use as the test stimuli. These steps varied in how often they were labeled as /s/ or /ʃ/, while avoiding floor and ceiling effects. Presentation during the test block was auditory only (following Kraljic et al., 2008).

2.1.3 Procedure.—The experiment began with instructions, followed by an exposure block, a test block, and a post-experimental survey. The exposure block manipulated the perceptually shifted sound and the type of causal attribution. All other blocks were identical across all the conditions.

At the beginning the experiment, participants were instructed to transcribe two English words. They could listen to each word as many times as necessary to set their volume to a comfortable level, which they were asked not to change during the experiment. During exposure, participants completed a lexical decision task. For each trial, a fixation cross first appeared on the screen, followed by a video of the talker producing a single word or nonce word (inter-trial interval = 1000ms). During test, participants classified 6 steps along the /asi/-/aʃi/ continuum as either /asi/ or /aʃi/. These steps were played in ten cycles (henceforth, trial bins), each consisting of a random ordering of the 6 steps. To the participant, this appeared as one continuous block of 60 stimuli. Participants indicated their responses using the 'X' and 'M' keys on their keyboard. All key bindings were counterbalanced across participants.

Finally, participants answered a questionnaire that assessed their ethnic background, audio quality and equipment, technical difficulties, and attention during the task (Appendix B).

2.1.4 Exclusions.—All experiments reported below use the same exclusion criteria (Table 1). We excluded participants on five different criteria, discussed in turn.

Lexical decision accuracy.: Participants with low lexical decision accuracy (<85% on the normal pronunciations) might not have been paying adequate attention during the exposure task. Mean accuracy across all experiments was 96.2% correct (SD=3.1).

Repeat Participants.: Participants were asked to complete the experiment only once. Subsequent runs from participants were excluded.

No Headphones.: In previous perception experiments conducted on the web (Burchill, Liu, & Jaeger, submitted; Liu, Xie, Weatherholtz, & Jaeger, in prep), we found that participants who reported to be wearing headphones tended to show better task performance. We thus asked participants to wear headphones, and excluded participants who reported otherwise.

Wrong Answer on Catch Question.: Our post-experiment questionnaire asked participants whether the exposure speaker was a male or female. This question was easily answerable to participants who paid attention during exposure.

Swapped Keys (Inverted Slope).: Occasionally participants confused response keys during the test block, as evidenced by classification curves with inverted category boundaries (i.e. the participant provided more /ʃ/ responses for the more /s/-like part of the continuum). This is unexpected under any theory of speech perception. We excluded those participants.

Across all exclusion criteria, 9.7% and 15.9% of participants were excluded in Experiments 1a and 1b, respectively. This exclusion rate is within the range of other web-based perceptual recalibration experiments (Kleinschmidt & Jaeger, 2012: 14%-25% exclusions) and comparable lab-based experiments (Norris et al., 2003: 6%; Kraljic & Samuel, 2011: 5% - 14%).

2.2 Results.

For ease of interpretation, we present the results separately for Experiment 1a (causally unambiguous exposure: 10CU) and Experiment 1b (causally ambiguous exposure: 10CA), before presenting the combined results. Throughout this paper, analyses employ Generalized Linear Mixed Models (Breslow & Clayton, 1993), with link functions reflecting the nature of the various dependent variables (Baayen, Davidson, & Bates, 2008; Jaeger, 2008). All analyses, unless explicated stated otherwise, included the maximal random effect structure justified by the design (by-participant random intercepts).

2.2.1 Experiment 1a.—Experiment 1a tested whether our shorter, web-based crowdsourcing experiment could be used to detect perceptual recalibration in the presence of causally unambiguous percepts (10CU). Following Kraljic et al. (2008), we first analyze the lexical decision responses during exposure to establish that the atypical, shifted word pronunciations were perceived as words, as intended. These analyses, presented in Appendix D, validate the present web-based paradigm: participants correctly identified 98.9% of the critical items with normal pronunciations, and 93.5% of the critical items with atypical pronunciations as words. This is comparable to previous studies (e.g., Kraljic et al. 2008: 97.6% for normal items and 94.9% for atypical items). Supplementary analyses of the lexical decision data from all our experiments are provided in Appendix D.

Next, we analyze whether exposure had an effect on the perceived category boundary during test. Following previous perceptual recalibration studies (Kraljic & Samuel, 2005, 2011; Kraljic et al., 2008; Norris et al., 2003), participants in the ʃ-Label condition should shift the category boundary towards /s/ and participants in the S-Label condition should shift the category boundary towards /ʃ/. Figure 4 shows that this was indeed the case in Experiment 1a.

We begin with an omnibus analysis of this shift that collapses responses across the six continuum steps. This analysis follows(following previous work including Kraljic & Samuel, 2011; Kraljic et al., 2008). Mixed logit regression, predicting categorization by Label condition (sum-coded: ʃ-Label = 1 vs. S-Label = −1), confirmed the perceptual recalibration effect found in previous work: more /ʃ/ responses were observed in the ʃ-Label condition than in the S-Label condition (= 1.14, $z = 4.4$, $p < 0.001$, Table 2).

In the remainder of this paper, we employ an alternative analysis that we consider more appropriate for our purpose. Omnibus analyses implicitly assume that the perceptual recalibration effect is *not affected by testing it.* However, there are reasons to doubt this assumption: if adaptation is indeed due to continuous distributional learning (as proposed in Kleinschmidt & Jaeger, 2015), then the perceptual input experienced during test trials might affect later test responses. Such “unlearning” may result in a reduced perceptual recalibration effect in the later vs. initial trials of the test (Scharenborg & Janse, 2013; for additional discussion, see Norris et al., 2003). Specifically, we hypothesized that exposure to a uniform distribution of /asi/-/aʃi/ test stimuli during the test block would partially undo the effects of perceptual recalibration.[4] Omnibus tests that are conducted over the entire test block would thus underestimate the actual perceptual recalibration effect immediately following exposure. For the present purpose, this would be particularly problematic, as some of the experiments we present below have as a goal to test whether the perceptual recalibration effect is *reduced* under certain conditions.

To address this, we first show that testing indeed affects later categorization such that perceptual recalibration effects reduce over the course of testing. Then, we introduce a novel analysis that allows us to assess the perceptual recalibration effect immediately following the exposure block, i.e. providing an estimate for the perceptual recalibration at the *beginning* of the test block.

Recall that the test block consisted of ten bins (trial bins) of the six continuum steps. Figure 5 plots the proportion of /ʃ/ responses transformed into empirical logits as a function of Trial Bin. As hypothesized, the perceptual recalibration effect (the difference between the S-Label and ʃ-Label condition) is much larger at the onset of testing (first trial bin), but then reduces as testing continues. Specifically, the proportion of /ʃ/ responses in both Label conditions converges towards 50% (an empirical logit of 0).

One possibility to estimate the recalibration effect at the onset of testing is to use the responses in the first trial bin, or to use designs with shorter test phases. Here we pursue a different approach, which draws on the entirety of test responses, making it more robust to variability in the data. We used a mixed logit model to predict /ʃ/ responses from Label (sum-coded: ʃ-Label = 1 vs. S-Label = −1), Trial Bin, and their interaction. Trial Bin was coded continuously with the first trial bin coded as 0. This way the estimated effect of Label represents the estimate of the recalibration effect during the first trial bin of testing. The results of this analysis are summarized in Table 3. The predictions of the mixed logit model, transformed into empirical logits, closely matched participants' responses across trial bins (Figure 5).

This analysis returned a significant effect of Label, with more /ʃ/ responses in the ʃ-Label condition than in the S-Label condition ($\hat{\beta} = 1.65$, $z = 6.2$, $p < 0.001$). This is the estimate of the Label effect at the first trial bin. In line with Figure 5, there was also a significant main effect of Trial Bin, such that the total proportion of /ʃ/ responses increased over trial bins ($\hat{\beta} = 0.10$, $z = 8.5$,/ $p < 0.001$), and a significant interaction between Label and Trial Bin ($\hat{\beta} = -0.10$, $z = -8.8$, $p < 0.001$), such that the difference in /ʃ/ responses between the Label conditions decreased as a function of Trial Bin.

This convergence between the two Label conditions towards 50/50 /s/ and /ʃ/ responses (= 0 logits) over trial bins was reliable in the analyses of all experiments reported below (Appendix F). Given the close match between the model's predictions and given that recalibration effects were indeed much larger at the onset of the test block, we continue to use this approach to investigate differences in recalibration effects in the remainder of the

[4] While perceptual recalibration has been found to be stable under certain conditions (Eisner & McQueen, 2006; Kraljic & Samuel, 2006), this stability has been demonstrated in the *absence of additional evidence from the same talker*. These findings thus leave open whether testing with auditory stimuli from the same talker—as done here and in most previous work—affects perceptual recalibration.

paper. Here and for all experiments analyzed below, Trial Bin is always coded with the first trial bin coded as 0, and Label is always sum-coded (ʃ-Label = 1 vs. S-Label = −1).

2.2.2 Experiment 1b.—

Now that we have established that the basic perceptual recalibration effect can be detected using our materials and web-based paradigm, we ask whether we can replicate the effect of causally ambiguous evidence (10CA). Providing an incidental cause, such as a pen in the mouth, for an atypical pronunciation should reduce or completely block perceptual recalibration (Kraljic et al., 2008). Recall that, in contrast to Kraljic et al. (2008), our experiment does not present listeners with causally unambiguous normal pronunciations during the second half of the exposure block, which would disambiguate the cause as characteristic of the talker. In our experiment, listeners are instead only provided with causally ambiguous atypical tokens.

In deriving predictions for this manipulation, it is important to keep in mind that a specific incidental cause (in this case, a pen in the mouth) might affect the articulation of different sounds asymmetrically. For instance, a pen might plausibly disrupt lip-rounding, which is involved in the articulation of /ʃ/, whereas no such lip-rounding is involved in the articulation of /s/ (Ladefoged & Maddieson, 1996, p. 148). This makes the pen in the mouth a more plausible cause for an atypical /ʃ/ than an atypical /s/.[5]

If listeners take such asymmetries into account, the pen in the mouth should be more likely to block perceptual recalibration for the ʃ-Label condition, but less so or not at all for the S-Label condition. Indeed, the lexical decision data from Experiment 1b supports this prediction. We found that the critical words containing /ʃ/ were more likely to be accepted as words than those containing /s/ (see Appendix D). This suggests that the pen in the mouth was seen to be a more plausible cause for the /ʃ/ word (as they continued to be accepted as words) than the /s/ words (as they were less likely to be accepted as words).

Previous experiments by Kraljic and colleagues (2008) found that adaptation occurred when atypical pronunciations were shown in absence of an incidental cause (causally unambiguous) but blocked in the presence of an incidental cause (causally ambiguous). We perform the novel analysis described in Experiment 1a on the combined data from Experiment 1a (10CU) and 1b (10CA) in order to assess the effect of causal ambiguity on adaptation. We then compare Experiments 1a and 1b against the baseline, obtained from a separate set of participants.

We performed a mixed logit regression (Table 4) to analyze the proportion of /ʃ/ responses as a function of continuous Trial Bin, Label, Cause (10CU = 1 vs. 10CA = −1), and their interactions. There was a significant main effect of Label, such that the ʃ-Label condition labeled more stimuli as /ʃ/ than the S-Label condition ($\hat{\beta} = 1.28$, $z = 7.0$, $p < 0.001$). There

[5] Previous work has observed asymmetries in the amount of perceptual recalibration in the *absence* of alternative incidental causes (Samuel, 2016; Zhang & Samuel, 2014). For example, Samuel (2016) observed a bigger shift in the category boundary after exposure to causally unambiguous ʃ-labeled words, compared S-labeled words (both relative to a baseline condition). Our present argument focuses on asymmetries in the extent to which an incidental cause—the pen—can block recalibration (Experiment 1b, compared to Experiment 1a), rather than asymmetries in the amount of recalibration (differences between the ʃ- and S-label condition in Experiment 1). With regard to the latter, we note though that we observe no striking asymmetry in the amount of recalibration (cf. Figure 4).

also was a significant main effect of Cause ($\hat{\beta} = 0.38$, $z = 2.1$, $p < 0.05$): the 10CU condition provided more /ʃ/ responses at the beginning of the test block than the 10CA condition.

Critically, this main effect was driven by a significant interaction between Cause and Label ($\hat{\beta} = 0.36$, $z = 2.0$, $p < 0.05$), such that there was a larger difference between ʃ- vs. S-Label in the 10CU condition than in the 10CA condition. This means that participants showed less perceptual recalibration when they were exposed to atypical pronunciations that were causally ambiguous than causally unambiguous. Simple effect analyses further revealed that this held in the ʃ-Label condition (ʃ-Label: $\hat{\beta} = 0.74$, $z = 2.9$, $p < 0.01$), but not in the S-Label condition ($p > 0.93$). This suggests that having the pen in the mouth affected the two contrasts unequally, serving to block perceptual recalibration for /ʃ/, but not /s/.[6]

Finally, we compared all four exposure conditions (Cause × Label) to the baseline, in which participants only saw normal pronunciations, all of which were causally unambiguous. This result is visualized in Figure 6. Using a mixed effects regression model, we predicted the proportion of /ʃ/ responses from Exposure (treatment coded: baseline = reference level), Label, Trial Bin, and their interactions. Critically, we found evidence for perceptual recalibration away from the baseline in all conditions (ps < 0.05), except for the causally ambiguous ʃ-Label condition (p > 0.52; see Table 5). This corroborates with our results from the previous regression analysis, where we identified blocking of perceptual recalibration in the causally ambiguous ʃ-Label condition.

2.3 Discussion.

Experiment 1a replicates previous perceptual recalibration studies: listeners who are exposed to atypical pronunciations from a novel talker adapt their categorization of subsequent input from that talker when no plausible incidental cause is provided (causally unambiguous). This suggests that listeners attribute the unexpected pronunciations to the talker, taking it to be characteristic of that talker. When the same atypical pronunciations are experienced in the presence of a plausible alternative cause (causally ambiguous, Experiment 1b), listeners adapt *less*. This constitutes a conceptual replication of Kraljic et al. (2008) and validates our web-based paradigm. Like the findings of Kraljic and colleagues, our findings are compatible with the hypothesis that listeners attribute the unexpected pronunciations to the plausible incidental cause, rather than the talker.

One intriguing difference between our results and those of Kraljic et al. (2008) is that while they found complete blocking of recalibration for both contrasts when they were paired with an incidental cause, we found blocking only for the ʃ-Label condition. If listeners altogether ignore atypical pronunciations in the presence of an alternative cause (as proposed in Kraljic & Samuel, 2011), we should not see perceptual recalibration after exposure to atypical percepts when they are causally ambiguous—contrary to our findings for the S-Label condition. Under the present proposal, however, the asymmetry observed in our experiment receives an explanation.

[6] For the present purpose the specific way in which responses converge towards 50/50 /s/ and /ʃ/ responses across Trial Bins (reflected in main effects of, and/or interactions with, Trial Bin) are not of interest. In the interest of brevity, any such effects are thus only discussed and visualized in Appendix F. The same holds for all other experiments reported below.

Critically, our experiment differs from Kraljic et al. (2008) in that participants never received disambiguating evidence after the causally ambiguous pronunciations. In our experiment, we thus see the consequences of listeners' beliefs about what constitutes an *a priori* plausible cause for an atypical /s/, compared to an atypical /ʃ/. In particular, the pen might plausibly be perceived as disrupting lip-rounding, which is not involved in the articulation of /s/ (Ladefoged & Maddieson, 1996, p. 148). Indeed, our analysis of the lexical decision data from exposure revealed that participants tended to identify more atypical pronunciations as words when they were labeled as /ʃ/ than /s/, but only when paired with the pen in the mouth. This is consistent with the hypothesis that participants considered the pen in the mouth a sufficiently likely cause for atypical /ʃ/ pronunciations, but less so for atypical /s/ pronunciations.

Why then did Kraljic et al. (2008) observe blocking for both Label conditions? In Kraljic et al. (2008), disambiguating evidence always followed causally ambiguous percepts: this disambiguating evidence always showed the talker producing normal percepts with the pen in the hand. Under the present proposal, this disambiguating evidence is expected to overwrite the uncertainty participants had about the true cause for the atypical pronunciations observed in the first half of exposure: regardless of how plausible participants initially considered the pen as a cause for the observed pronunciations, they now receive evidence that the talker sounds normal.

There are also other differences between the present experiment and Kraljic et al. (2008) that might have caused the difference in results. This includes small differences in the procedure (web-based vs. lab-based), the length of the experiment, and, finally, the materials: although our auditory stimuli for the exposure block were identical to those used by Kraljic and colleagues, they were aligned with new videos (since the old videos from Kraljic et al., 2008 are no longer available). These videos showed a different talker and (likely) different pen placements. As a consequence, it is possible that the alignment between the auditory stimulus and the video might differ between the two studies. Any of these reasons might theoretically have caused the differences in results. We note, however, that—for this to be the case—these differences would have to selectively affect participants' interpretation of causally ambiguous /ʃ/ pronunciations in our experiment: our Experiment 1 replicated Kraljic et al. (2008) for causally unambiguous /s/- and /ʃ/-labeled exposure and for causally ambiguous /s/-labeled exposure.

As we will see below, the results of Experiments 2 and 3 indeed favor the first explanation we offer above—when listeners receive only causally ambiguous evidence for a talker's pronunciations, they can maintain uncertainty about talker-specific characteristics.

3 Experiment 2a and 2b

In Experiment 2, we begin to address how the listener's uncertainty about the cause of a typical pronunciation may affect their subsequent adaptation. We ask whether listeners ignore causally ambiguous percepts, or if they maintain perceptual information about causally ambiguous percepts. If the latter is the case, then listeners may be able to disambiguate the cause of previously experienced causally ambiguous percepts, given additional causally unambiguous evidence.

Using the same paradigm as in Experiment 1, Experiment 2 exposes participants to causally ambiguous or unambiguous atypical pronunciations (between participants), and then to causally unambiguous atypical pronunciations to disambiguate previously experienced causally ambiguous pronunciations. If listeners are sensitive to the causal uncertainty of atypical pronunciations, we should replicate previous work (Kraljic et al., 2008; Kraljic & Samuel, 2011): causally disambiguating evidence following causally ambiguous evidence should lead listeners to adapt. Critically, if listeners can maintain (some aspects of) causally ambiguous percepts over many trials in the exposure block, the *degree* of perceptual recalibration listeners exhibit should be greater than what is expected based on solely the number of disambiguating percepts. Further, if causally ambiguous evidence is maintained *perfectly* for the duration of the experiment, even a few disambiguating percepts with atypical pronunciations should lead listeners to exhibit perceptual recalibration proportional to the total number of atypical pronunciations (i.e., the combined number of causally ambiguous and causally unambiguous atypical trials).

3.1 Methods.

3.1.1 Participants.—We recruited 238 participants for Experiment 2a, and 108 participants for Experiment 2b on Amazon Mechanical Turk (aiming to hold constant, the number of participants per between-participant condition across experiments). Participants were paid $1.80 (Experiment 2a: 160 exposure trials) or $1.20 (Experiment 2b: 60 exposure trials) for their time ($6/hour). Test trials and recruitment criteria were identical to Experiment 1.

3.1.2 Materials.—In addition to the stimuli we used in Experiment 1a and 1b, we incorporated an additional 60 words (12 critical and 48 filler words) from Kraljic et al. (2008) to each of the four lists (words given in Appendix A). These critical words were always paired with videos of the speaker with a pen in her hand, which provides causally unambiguous evidence that the shifted production is characteristic of the talker.

3.1.3 Procedure and Exclusions.—The procedure and instructions for Experiment 2a were identical to Experiment 1a and 1b, with the exception that the exposure block was longer (see Figure 7). One group of participants experienced 16 causally unambiguous atypical pronunciations, while the other experienced 10 causally ambiguous atypical pronunciations, followed by 6 causally unambiguous atypical pronunciations. For the second group of participants, these final 6 pronunciations disambiguated the cause of the prior 10 atypical pronunciations as being characteristic of the speaker. Henceforth, these conditions will be referred to as the 16CU and 10CA+6CU conditions, respectively.

3.2 Results

Following Experiment 1a and 1b, we analyze the categorization responses from the test block in Experiment 2a to test whether participants maintain uncertainty about the cause of atypical pronunciations. We then analyze the combined data from Experiments 2a and 2b to assess the possibility that the results of Experiment 2a may be partially driven by a ceiling effect.

3.2.1 Experiment 2a.—

The purpose of this experiment was to better understand how listeners treat evidence from causally ambiguous pronunciations: do listeners ignore it or encode it separately? Or do listeners maintain perceptual information for these causally ambiguous percepts, along with the information that these percepts were causally ambiguous, in a way that would enable them to benefit from this prior information if later given causally disambiguating evidence? If the latter is the case and if participants are able to *perfectly* maintain information about previous percepts, then we would except no difference in perceptual recalibration between the 16CU and 10CA+6CU conditions: participants who are exposed to 10 causally ambiguous items followed by 6 disambiguating causally unambiguous items should adapt as if they had seen 16 causally unambiguous items. However, if the former were the case, then we would predict less perceptual recalibration in the 10CA+6CU condition (as it contains 10 causally ambiguous items that would be ignored, leaving only 6 causally unambiguous items to drive recalibration) than in the 16CU condition (as it contains 16 causally unambiguous items that would drive recalibration). This would also be consistent with *imperfect* maintenance of information, requiring an additional experiment (explained below) to clarify the results.

We conducted the same mixed logit regression as in Experiment 1, predicting proportion of /ʃ/ responses as a function of Label, Cause (sum-coded: 16CU = 1 vs. 10CA+6CU = −1), Trial Bin, and their interactions. The results are given in Table 6. Replicating Experiment 1, we found a significant main effect of Label, such that participants in the ʃ-Label condition categorized more sounds as /ʃ/ than participants in the S-Label condition ($\hat{\beta}$= 1.61, z = 8.6, p < 0.001). Crucially, we find neither a significant effect of Cause, nor an interaction between Label and Cause (all ps > 0.4). This suggests that participants who had seen 10 causally ambiguous tokens followed by 6 causally unambiguous tokens exhibited perceptual recalibration that is indistinguishable from those who had seen 16 causally unambiguous tokens.

3.2.2 Experiment 2b.—

The results of Experiment 2a is predicted by the proposal that listeners maintain (rather than ignore) causally ambiguous evidence for atypical pronunciations, and are able to integrate this casually ambiguous evidence with later occurring causally disambiguating evidence. Specifically, the lack of any difference between the 16CU and 10CA+6CU conditions of Experiment 2a seems to suggest that participants are able to perfectly maintain information about causally ambiguous percepts over the course of the experiment. This result is, however, also compatible with an alternative explanation: participants may rely only on the most recent causally unambiguous percepts. If participants rely on six or fewer percepts, then the two exposure conditions of Experiment 2a are expected to yield the same results (as they do). This explanation encompasses the possibility that participants may only *imperfectly* maintain information over the course of the experiment, thus relying on (e.g.) the most recent percepts. To address this alternative explanation, we conducted Experiment 2b. We exposed participants only to the final exposure block from Experiment 2a, which contained six causally unambiguous atypical pronunciations (6CU condition; Figure 7). If we find a smaller degree of perceptual recalibration in the 6CU condition, compared to the 10CA+6CU and 16CU conditions, this would unambiguously suggest that participants indeed maintain information about causally ambiguous pronunciations in Experiment 2a. If, on the other hand, we find identical degrees of perceptual recalibration in the 6CU condition and the 10CA+6CU and 16CU conditions, this would leave the interpretation of Experiment 2 ambiguous between the two explanations offered above.

We compare adaptation following exposure to 6 causally unambiguous percepts (6CU) to perceptual recalibration following the exposure conditions from Experiment 2a (16CU and 10CA+6CU; Figure 7). We extended the mixed logit regression for Experiment 2a, predicting proportion of /ʃ/ responses as a function of Cause (sum-coded: 16CU = 1 vs. 6CU = −1 and 10CA+6CU = 1 vs. 6CU = −1), Label, Trial Bin, and their interactions. The results of this regression are shown in Table 7.

We identified a significant effect of Label ($\hat{\beta} = 1.4$, $z = 9.3$, $p < 0.001$), such that the proportion of /ʃ/ responses was higher for the ʃ-Label group than the S-Label group at the first trial bin across Cause conditions. This is consistent with adaptation. Furthermore, the 6CU condition did not differ significantly from the 16CU or 10CA+6CU conditions in the proportion of /ʃ/ responses provided by the ʃ- vs. S-Label groups (*p*s > 0.6). This suggests that there was adaptation even after exposure to only six causally unambiguous percepts. This effect is shown in Figure 8 (rightmost errorbars). As also shown in Figure 8, adaptation (the effect of Label at the start of test) was numerically larger in the 10CA+6CU condition than in the 6CU condition. This would suggest that participants maintained causally ambiguous evidence, and later integrated this evidence after receiving causally disambiguating evidence. However, the critical interaction between Label and the comparison between the 10CA+6CU and 6CU conditions was not significant ($p = 0.8$), leaving open the possibility that the results from the 10CA+6CU condition were exclusively driven by the last final 6CU percepts.

3.3 Discussion

In Experiment 2a, we find that listeners exhibit perceptual recalibration when causally ambiguous percepts are followed by causally disambiguating percepts. Critically, we find that the degree of exposure seems to be identical for both between-participant conditions of Experiment 2a: we observe the same shift in categorization boundaries after causally ambiguous pronunciations are followed by causally unambiguous pronunciations, compared to the same total number of causally unambiguous pronunciations. This result is expected under the proposal that speech adaptation can be understood as causal inference under uncertainty. More specifically, it would seem to suggest that participants maintain information about the causally ambiguous pronunciations perfectly for the duration of the exposure block.

However, at the same time, the comparison between Experiments 2a and 2b provides reason for caution: we cannot reject the hypothesis that the last six critical items of the 10CA+6CU condition are sufficient to cause perceptual recalibration that is (statistically) indistinguishable from recalibration after 10 or 16 causally unambiguous items. In essence, Experiment 2a might suffer from a ceiling effect. Indeed, additional analyses reported in Appendix E failed to rule out this possibility. The results of Experiment 2 are thus not conclusive: while Figure 8 reveals a pattern of results that would be rather unexpected under the hypothesis that listeners simply ignore causally ambiguous percepts, Experiment 2 by itself does not provide sufficiently strong evidence against this hypothesis.

4 Experiment 3

In Experiment 3, we again address the question of whether participants may be ignoring or maintaining information about causally ambiguous percepts. To do this, we again present participants with causally ambiguous atypical pronunciations followed by disambiguating causally unambiguous pronunciations. However, in order to avoid problems with a potential ceiling effect as in Experiment 2, we shorten the exposure block (while maintaining the same critical-to-filler ratio as in Experiments 1 and 2). We compare perceptual recalibration following exposure to 8 causally ambiguous percepts followed by 2 causally unambiguous percepts (8CA+2CU) with exposure to 10 causally unambiguous percepts (10CU). Paralleling Experiment 2, we also compare this to exposure to only 2 causally unambiguous percepts (2CU).

Under the hypothesis that causally ambiguous percepts are ignored or encoded separately, we would expect that perceptual recalibration n in the 8CA+2CU condition – where the first 8 percepts are effectively ignored – would be *identical* to the 2CU condition. However, if listeners are able to maintain (even some limited degree of) perceptual information about the previously experienced percepts, then we would expect stronger recalibration in the 8CA +2CU condition than in the 2CU condition. Whether or not the previous 8 percepts are perfectly maintained is addressed by the addition comparison of the 10CU condition and the 8CA+2CU condition. If listeners are able to *perfectly* maintain the information from the initial 8 causally ambiguous percepts, then we expect identical degrees of recalibration in the 8CA+2CU condition compared to 10CU condition. However, if listeners are able to maintain limited information, then we expect less recalibration in the 8CA+2CU condition, compared to the 10CU condition.

4.1 Methods.

4.1.1 Participants.—We recruited 118 participants for the 8CA+2CU condition and 113 participants for the 2CU condition on Amazon Mechanical Turk. Participants in the 8CA +2CU condition were paid $1.50 (100 exposure trials), and participants in the 2CU condition (20 exposure trials) were paid $0.50 ($6/hour).

4.1.2 Materials.—The materials we used in the 8CA+2CU condition were identical to those used in Experiment 1. In both experiments, participants heard 10 items containing atypical pronunciations of one fricative, 10 items containing normal pronunciations of the other fricative, and 80 filler items. The only difference was that while these atypical critical items were always presented with causally unambiguous videos (pen-in-hand) in Experiment 1a, the first 8 items were presented causally ambiguous videos (pen-in-mouth) in the 8CA +2CU condition of Experiment 3. The final 2 items in the 8CA+2CU condition were shown with causally unambiguous videos, thus disambiguating the cause of the prior 8 atypical pronunciations as being characteristic of the speaker. This is illustrated in Figure 7. The materials we used in the 2CU condition consisted of the final block of 20 items from Experiment 1a.

4.1.3 Procedure and Exclusions.—The exact instructions, procedure, and exclusions were identical to Experiments 1 and 2 (for exclusions, see Table 1).

4.2 Results

The purpose of this experiment was to better understand how participants treat causally ambiguous evidence. We compare categorization data from the test across the different exposure conditions to test our predictions about how causally ambiguous information is treated. To do this, we compare the 8CA+2CU condition with the 2CU condition, and compare the 8CA+2CU condition with the 10CU condition from Experiment 1a.

We conducted a mixed logit regression to compare the strength of perceptual recalibration in 8CA+2CU, 2CU, and 10CU conditions. This regression predicted proportion of /ʃ/ responses from Label, Cause (sum coded: 2CU = 1 vs. 8CA+2CU = −1 and 10CU = 1 vs. 8CA+2CU = −1), Trial Bin, and their interactions. The output of the regression is given in Table 8.

We identified a significant effect of Label ($\hat{\beta} = 1.05$, $z = 6.6$, $p < 0.001$), such that the proportion of /ʃ/ responses was higher for the ʃ-Label group than the S-Label group at the first trial bin. This is in the expected direction. Additionally, we find that there were significantly more /ʃ/ responses in the 2CU condition, compared to the 8CA+2CU condition ($\hat{\beta} = 0.53$, z =2.4, $p < 0.05$). Crucially, this was driven by a significant interaction with Cause for the 2CU condition vs. 8CA+2CU condition ($\hat{\beta} = -0.65$, $z = -2.9$, $p < 0.01$), such that the difference in proportion of /ʃ/ responses between Label conditions was smaller for the 2CU condition than the 8CA+2CU condition. This is consistent with there being less adaptation in the 2CU condition than in the 8CA+2CU condition. Crucially, it suggests that the 8 causally ambiguous percepts were not ignored by participants (Figure 9).

Additionally, we identified that a significant interaction with Cause for the 10CU condition vs. 8CA+2CU condition ($\hat{\beta} = 0.62$, $z = 2.74$, $p < 0.01$), such that the difference in proportion of /ʃ/ responses between Label conditions was larger for the 10CU condition than the 8CA +2CU condition. This points to limited, rather than perfect, maintenance of causally ambiguous evidence. Simple effects analysis assessing the effect of Label on each of the Cause conditions confirmed that there was a significant perceptual recalibration effect in the 10CU ($\hat{\beta} = 1.67$, $z = 6.0$, $p < 0.001$) and 8CA+2CU ($\hat{\beta} = 1.09$, $z = 4.01$, $p < 0.001$) conditions, but not in the 2CU condition ($\hat{\beta} = 0.4$, $z = 1.44$, $p > 0.14$). Taken together, these results suggest that listeners are able to capable of limited maintenance of causally ambiguous evidence, such that they are able to use this evidence to adapt when later provided with disambiguating causally unambiguous evidence.

4.3 Discussion.

Paralleling the results from Experiment 2a and 2b, the results from Experiment 3 provide further support for the hypothesis that listeners engage in causal reasoning during speech perception and do not ignore causally ambiguous percepts. These results are compatible with the idea of limited, rather than perfect, maintenance of causally ambiguous percepts: participants can use previously experienced causally ambiguous evidence to adapt after having been provided with causally disambiguating evidence.

5 General Discussion

In this paper, we tested the hypothesis that adaptation to atypical pronunciations during speech perception involves causal inferences, and that listeners can maintain evidence from causally ambiguous pronunciations. For instance, an atypical pronunciation experienced in the presence of a plausible incidental cause (i.e., pen in the mouth) may later be shown to have arisen from a characteristic cause (i.e., lisp). When a listener is uncertain of the true cause behind a pronunciation, it may be beneficial for the listener to maintain, rather than to ignore, information about causally ambiguous percepts, such that that they could potentially draw from this information later on, given disambiguating information.

Our results suggest that listeners do not indiscriminately adapt to both causally ambiguous and causally unambiguous evidence: in Experiment 1, we showed that perceptual recalibration is blocked when listeners experience causally ambiguous (in line with Kraljic et al., 2008). In Experiment 2 and 3, we addressed the question of whether such blocking occurs because listeners *ignore* causally ambiguous percepts or if listeners maintain information about causally ambiguous percepts. We find support for the latter: the degree of perceptual recalibration following exposure to a combination of causally ambiguous and causally unambiguous percepts was not identical to the degree of perceptual recalibration following exposure to only the causally unambiguous percepts alone. We find that listener behavior is consistent with a view under which listeners are capable of limited, rather than perfect, maintenance of perceptual evidence from causally ambiguous percepts, and are able to update their beliefs following exposure to causally disambiguating evidence.

Before we discuss these results further, we note that they were obtained via Amazon Mechanical Turk, a web-based crowdsourcing platform. There have been other web-based studies on speech perception (Burchill et al., submitted; Byun, Halpin, & Szeredi, 2015; Kleinschmidt & Jaeger, 2012; Kleinschmidt, Raizada, & Jaeger, 2015; Kunath & Weinberger, 2010; Liu et al., in prep; Xie et al., in prep), but the present study is, to the authors' knowledge, the first web-based perceptual recalibration study of fricatives. This is of note because fricative identification relies on high-frequency noise information (Scharenborg, Weber, & Janse, 2015), which may be transmitted more or less faithfully by different audio equipment or listening conditions. The successful detection of perceptual recalibration across our experiments suggests that web-based experiments can (at least to some extent) be used even when studying coarse-grained effects on the perception of sound contrasts distinguished by spectral features (incl. fricatives and vowels).

In the remainder of this paper, we first discuss why it may be beneficial for listeners to engage in causal reasoning during speech perception. We then situate our results in the context of other work on the maintenance of perceptual information during speech processing, and other work on causal inferences in language processing.

5.1 Deciding what previous experience to draw on for categorization

How do listeners know *when* to adapt? At first blush, it might seem tempting to assume that the speech perception system is continuously and indiscriminately adapting to any input it receives. Indeed, some models of implicit learning during language processing predict just that (e.g., Chang et al., 2006; Dell & Chang, 2014; Reitter et al., 2011). However, as we have outlined in the introduction, there are a priori considerations why an effective speech perception system *should not* indiscriminately adapt in this manner—namely, adapting indiscriminately without consideration of causes risks unnecessarily volatile category recognition (Kleinschmidt & Jaeger, 2015; Samuel, 2011). Adapting to an atypical pronunciation that arose from an incidental cause would not aid the listener in developing an accurate model of how a talker characteristically sounds.

Rather than framing the problem in terms of *when listeners adapt*, we can frame the problem in terms of inferring *what previous experience listeners draw upon* when interpreting the current speech signal. At any given point in time, there can be many reasons why talkers sound as they do. The listener's inference problem then is to determine which stored representations (e.g., in the senses of exemplars or episodes: Johnson, 1997; Pierrehumbert, 2002, or in the sense of talker-, group- or situation-specific generative models: Kleinschmidt & Jaeger, 2015) to use to interpret the present speech signal. For example, in the present experiments, participants listening to the auditory-only test trials need to decide which previous experience, or mixture of previous experiences, to use to categorize sounds into /s/ and /ʃ/. Similarly, when causally ambiguous percepts are paired with normal pronunciations (as in Kraljic & Samuel, 2011: Experiment 3), listeners need to decide the relative probabilities to assign to the possibility that the pen did not affect pronunciation at all or if somehow the pen resulted in normalizing a talker's otherwise atypical pronunciation. Additional evidence for one possibility or the other can lead participants to adjust which previous experiences they draw upon during perception.

This alternative perspective thus asks how listeners determine what previous experiences are relevant to for the recognition of the current input. That is, the causal inferences we refer to directly affect the recognition of phonemes and words (i.e., language understanding). The framing we have employed in the introduction, on the other hand, asks how listeners determine what the current input tells them about future inputs from the same talker. In this perspective, the causal inferences affect whether a percept is stored or how it is integrated into existing representations (i.e., adaptation).

Although the present study does not allow us to distinguish between these two—mutually compatible—perspectives, they differ in important ways. The two views differ, for example, in whether they allocate the complexity of causal inference to online language understanding or to the adaptive learning processes that accompany language understanding. The two views also make different assumptions about the information encapsulation of memory processes—indiscriminate automatic storage of all percepts (as in standard episodic and exemplar models:,Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2002) vs. ‘smart’ integration of percepts into structured representations, guided by causal inferences. In the latter case, for example, listeners might learn and represent models of how talkers sound in particular circumstances. For example, it might be beneficial for listeners to have a representation of how a particular talker (or talkers in general) sound when intoxicated or when talking with their mouth full (see also Kleinschmidt & Jaeger, 2015, p. 184). This is particularly the case with causes that occur frequently either across talkers or within the same talker (i.e. imagine a talker who has a strong tendency to talk with food in their mouth).

For all these reasons, we consider these two perspectives an interesting venue for future research. Regardless which of these two perspectives we take, however, causal inference provides listeners with a way to take into account *why* a talker might sound as they do. Under uncertainty about whether an atypical pronunciation is characteristic of or incidental to the talker, it may be beneficial for the listener to maintain information about this percept, in case they receive evidence later on that can disambiguate this prior information. This would enable the listener to draw upon all of the prior evidence received from a talker to facilitate current and future language understanding.

5.2 Limits of maintaining previously experienced percepts

Our experiments also point to the limits of this system. While we found that listeners can maintain information about audio-visual percepts for a surprisingly long time (over many trials), listeners’ maintenance of uncertainty was not perfect: after disambiguation, previously experienced causally ambiguous percepts contribute to the overall perceptual recalibration effect, but not to the same extent that previously experienced causally unambiguous percepts do.

Our experiments concern the maintenance of causal information over the course of an experiment, and relates to (but should be distinguished from) recent work on right context effects in speech perception. The growing body of work on right context effects reveals that listeners may maintain uncertainty and update their beliefs about earlier speech segments based on information from further downstream in the speech signal (for a recent review, see Dahan, 2010). For instance, Connine, Blasko, and Hall investigated the temporal constraints on ambiguity resolution by presenting participants with sentences containing perceptual/lexical ambiguities that were disambiguated downstream in the sentence. They found that retroactive interpretations occurred for disambiguating information that occurred three syllables downstream, but not 6-8 syllables downstream, pointing to a potential limit of this system (Connine et al., 1991). However, Bicknell, Tanenhaus, and Jaeger (2015), after addressing a methodological concern in the original study, found evidence for maintenance even at 6-8 syllables, the longest delay tested (see also Bushong & Jaeger, 2017). This would seem to suggest perfect or near perfect maintenance of information, at least at the 6-8 syllable maximum delays tested.

However, compared to the work we present here, such experiments have only tested maintenance on the order of syllables (seconds), rather than over the course of an entire experiment (minutes) as we do here. This difference in timescale could be one potential reason for why we find limited, rather than perfect, maintenance in our experiments.

One way that listeners may accomplish this is by encoding the auditory percept along with the (visual) context, using the combined episode during later speech perception. This idea is compatible with episodic accounts and exemplar-based accounts of speech perception that assume rich storage of information (Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2002), i.e. episodic trades of each percept experienced (see also Szostak & Pitt, 2013). An open question is whether other relevant information that influences the listener's uncertainty about the cause of the pronunciation, such as how likely a given incidental cause is to affect a particular speech sound, is stored alongside these episodes, or only inferred later on during the categorization process.

Such storage of perceptual information is not unheard of. There is evidence for detailed storage of visual and auditory evidence. For example, participants have been found to be able to successfully maintain detailed (though, of course, not error-free) information for thousands of novel images (Brady, Konkle, Alvarez, & Oliva, 2008). Additionally, participants have been found to maintain talker-specific perceptual adjustments for at least 12 hours after exposure (Eisner & McQueen, 2006; Vroomen & Baart, 2009; Witteman, Bardhan, Weber, & McQueen, 2015; Xie & Myers, 2016).

Like the aforementioned experiments on right-context effects, along with the vast majority of experiments on perceptual recalibration, our experiments involve the manipulation of a very small set of sound contrasts (/s/ and /ʃ/). For each participant, all atypical productions in our experiments were limited to (multiple instances of) one phoneme. Consequently participants only need to maintain information relevant to one atypical sound contrast. It is unclear from our experiment whether participants would be able to similarly maintain causal uncertainty under the demands of everyday speech perception, i.e. where there are a broader set of potential causes and a larger set of speech sounds, which might show influence from those causes (for discussion, see also Burchill et al., submitted).

5.3 Causal inferences during language processing.

The present study also contributes to work on causal attribution during language processing, extending it to the domain of speech perception. Previously, the role of causal inferences has been investigated with regards to both reference resolution and pragmatic interpretation. In a visual-world eye-tracking experiment, Arnold et al. (2007) found that participants directed more eye movements towards unfamiliar objects when the instructions included a disfluency (e.g., *"Click on [pause] thee uh red … "*), revealing that listeners expected the talker to refer to an unfamiliar object following a disfluency. However, these expectations were sharply reduced when participants were provided with an alternative plausible cause for the talker's disfluency (participants were told that the speaker had object agnosia and difficulty naming familiar objects). This suggests that online reference resolution can draw on top-down causal inferences about the speaker's cognitive state. In another eye-tracking study, Grodner & Sedivy (2007) found that pragmatic inferences that typically are observed were cancelled when participants were told that the talker "an impairment that caused language and social problems”. In both of these studies, sensitivity to specific linguistic cues is disrupted when the presence of those cues can be plausibly attributed to alternative causes.

The present study differs from these works in that the studies of both Arnold et al. (2007) and Grodner & Sedivy (2011) involve explicit top-down provision of alternative plausible causes. By contrast, participants in the present study *were not* verbally informed of any impairment of the talker or potential role of the pen in the talker’s productions. This suggests that causal inferences in language processing can also arise spontaneously without explicit instructions. Additionally, it suggests that such causal inferences may draw upon a diverse array of sources, including visual context and speaker knowledge.

Finally, one aspect of our results provides tentative evidence that listeners may be highly attuned to specific causes and the result they have on specific sound contrasts. In Experiment 1b, we found that while having a pen in the mouth served to block perceptual recalibration for the /ʃ/ sound, it did not have an effect on the /s/ sound. We hypothesized that this could be because a pen might plausibly disrupt lip-rounding, which is involved in the articulation of /ʃ/ but not /s/. This would mean that listeners have some degree of sensitivity to the articulatory gestures used in the production of specific sounds. Further experiments that address this possibility more directly strike us as an interesting opportunity for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Arty Samuel for generously providing the audio stimuli and for his helpful feedback on earlier presentations of this work. We also thank Molly Babel and Jamie Russell for generously providing the video stimuli used in this experiment and Dave Kleinschmidt for his help setting up the web-based experiment, and members of both the Human Language Processing and KurTan lab at Rochester as well as the Prosody Lab at McGill for feedback on earlier presentations of this work. Some of the results were presented at the 2015 Architectures and Mechanism of Language Processing Conference. This work was funded by the NIH R01 grant HD075797 to T. Florian Jaeger. The views expressed here are those of the authors and not necessarily those of the funding agencies.

References

Babel M (2016). Replication of Kraljic T, Samuel AG, Brennan SE (2008, PS 19(4)). Retrieved from osf.io/pj5hb

Arnold JE, Kam CLH, & Tanenhaus MK (2007). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(5), 914–930.

Baayen RH, Davidson DJ, & Bates DM (2008). Mixed-effects modeling with crossed random effects. Journal of Memory and Language, 59(4), 390–412.

Baese-Berk MM, Bradlow AR, & Wright BA (2013). Accent-independent adaptation to foreign accented speech. The Journal of the Acoustical Society of America, 133(3), EL174–EL180. [PubMed: 23464125]

Bicknell K, Tanenhaus MK, & Jaeger TF (2015). Listeners can maintain and rationally update uncertainty about prior words. Submitted for publication.

Bradlow AR, & Bent T (2008). Perceptual adaptation to non-native speech. Cognition, 106(2), 707–729. [PubMed: 17532315]

Brady TF, Konkle T, Alvarez GA, & Oliva A (2008). Visual long-term memory has a massive storage capacity for object details. Proc Natl Acad Sci U S A, 105(38), 14325–14329. doi:10.1073/pnas.0803390105 [PubMed: 18787113]

Breslow N, & Clayton D (1993). Approximate Inference in Generalized Linear Mixed Models. Journal of the American statistical Association, 88(421), 9–25.

Burchill Z, Liu L, & Jaeger TF (submitted). Maintaining perceptual information during accent adaptation.

Bushong W, & Jaeger TF (2017). Maintenance of perceptual information in speech perception. Paper presented at the Thirty-Ninth Annual Conference of the Cognitive Science Society.

Byun TM, Halpin PF, & Szeredi D (2015). Online crowdsourcing for efficient rating of speech: A validation study. Journal of communication disorders, 53, 70–83. [PubMed: 25578293]

Chang F, Dell GS, & Bock K (2006). Becoming syntactic. Psychological review, 113(2), 234. [PubMed: 16637761]

Chin SB, & Pisoni DB (1997). Alcohol and Speech: Academic Press.

Clarke CM, & Garrett MF (2004). Rapid adaptation to foreign-accented English. The Journal of the Acoustical Society of America, 116(6), 3647–3658. [PubMed: 15658715]

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

Dell GS, & Chang F (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philos Trans R Soc Lond B Biol Sci, 369(1634), 20120394. doi:10.1098/rstb.2012.0394 [PubMed: 24324238]

Eisner F, & McQueen JM (2006). Perceptual learning in speech: stability over time. J Acoust Soc Am, 119(4), 1950–1953. [PubMed: 16642808]

Goldinger SD (1998). Echoes of echoes? An episodic theory of lexical access. Psychological review, 105(2), 279.

Jaeger TF (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434–446. [PubMed: 19884961]

Johnson K (1997). Speech perception without speaker normalization: An exemplar model In Johnson K & Mullennix J (Eds.), Talker Variability in Speech Processing (pp. 145–165). San Diego, CA: Academic Press.

Johnson K, Pisoni DB, & Bernacki RH (1990). Do voice recordings reveal whether a person is intoxicated? A case study. Phonetica, 47(3-4), 215–237. [PubMed: 2130381]

Klatt DH (1986). The problem of variability in speech recognition and in models of speech perception. Invariance and variability in speech processes, 300–319.

Kleinschmidt DF, & Jaeger TF (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. Paper presented at the ACL Workshop on Cognitive Modeling and Computational Linguistics.

Kleinschmidt DF, & Jaeger TF (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. Paper presented at the Annual Conference of the Cognitive Science Society, Sapporo, Japan.

Kleinschmidt DF, & Jaeger TF (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. Psychological Review, 122(2), 148–203. [PubMed: 25844873]

Kleinschmidt DF, Raizada R, & Jaeger TF (2015). Supervised and unsupervised learning in phonetic adaptation. Paper presented at the CogSci.

Kraljic T, & Samuel AG (2005). Perceptual learning for speech: Is there a return to normal? Cognitive psychology, 51, 141–178. [PubMed: 16095588]

Kraljic T, & Samuel AG (2011). Perceptual learning evidence for contextually-specific representations. Cognition, 121, 459–465. [PubMed: 21939965]

Kraljic T, Samuel AG, & Brennan SE (2008). First impressions and last resorts how listeners adjust to speaker variability. Psychological science, 19, 332–338. [PubMed: 18399885]

Kunath SA, & Weinberger SH (2010). The wisdom of the crowd's ear: speech accent rating and annotation with Amazon Mechanical Turk. Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

Ladefoged P, & Maddieson I (1996). The sounds of the world's languages: Wiley-Blackwell.

Lancia L, & Winter B (2013). The interaction between competition, learning, and habituation dynamics in speech perception. Laboratory Phonology, 4(1), 221–257.

Liu L, Xie X, Weatherholtz K, & Jaeger TF (in prep). Adaptation and generalization to foreign-accented speech.

McMurray B, Rhone A, & Galle M (2012). FricativeMakerPro.

Norris D, McQueen JM, & Cutler A (2003). Perceptual learning in speech. Cognitive psychology, 47(2), 204–238. [PubMed: 12948518]

Nygaard LC, Sommers MS, & Pisoni DB (1994). Speech perception as a talker-contingent process. Psychological Science, 5(1), 42–46. [PubMed: 21526138]

Pardo JS, & Remez RE (2006). The perception of speech In Traxler M & Gernsbacher M (Eds.), The handbook of psycholinguistics (2 ed., pp. 201–248). New York, New York: Academic Press.

Pierrehumbert J (2002). Word-specific phonetics. Laboratory Phonology, 7, 101–139.

Pisoni DB, & Martin CS (1989). Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses. Alcohol Clin Exp Res, 13(4), 577–587. [PubMed: 2679214]

Qian T, Jaeger TF, & Aslin RN (2012). Learning to represent a multi-context environment more than detecting changes. Front Psychol, 3, 228. doi: 10.3389/fpsyg.2012.00228 [PubMed: 22833727]

Reitter D, Keller F, & Moore JD (2011). A computational cognitive model of syntactic priming. Cogn Sci, 35(4), 587–637. doi:10.1111/j.1551-6709.2010.01165.x [PubMed: 21564266]

Samuel AG (2011). The Lexicon and Phonetic Categories: Change is Bad, Change is Necessary In Gaskell GM & Zwitserlood P (Eds.), Lexical Representation: A Multidisciplinary Approach: Walter de Gruyter.

Samuel AG (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. Cognitive psychology, 88, 88–114. [PubMed: 27423485]

Scharenborg O, & Janse E (2013). Comparing lexically guided perceptual learning in younger and older listeners. Atten Percept Psychophys, 75(3), 525–536. doi:10.3758/s13414-013-0422-4 [PubMed: 23354594]

Scharenborg O, Mitterer H, & McQueen JM (2011). Perceptual learning of liquids. Paper presented at the Interspeech, Florence, Italy.

Scharenborg O, Weber A, & Janse E (2015). Age and hearing loss and the use of acoustic cues in fricative categorization. The Journal of the Acoustical Society of America, 138(3), 1408–1417. doi: 10.1121/1.4927728 [PubMed: 26428779]

Sidaras SK, Alexander JE, & Nygaard LC (2009). Perceptual learning of systematic variation in Spanish-accented speech. J Acoust Soc Am, 125(5), 3306–3316. doi: 10.1121/1.3101452 [PubMed: 19425672]

Sobin C, & Alpert M (1999). Emotion in speech: the acoustic attributes of fear, anger, sadness, and joy. J Psycholinguist Res, 28(4), 347–365. [PubMed: 10380660]

Szostak CM, & Pitt MA (2013). The prolonged influence of subsequent context on spoken word recognition. Atten Percept Psychophys, 75(7), 1533–1546. doi:10.3758/s13414-013-0492-3 [PubMed: 23801323]

Vroomen J, & Baart M (2009). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. Language and speech, 52(2-3), 341–350. [PubMed: 19624035]

Vroomen J, van Linden S, De Gelder B, & Bertelson P (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. Neuropsychologia, 45(3), 572–577. [PubMed: 16530233]

Weatherholtz K, & Jaeger TF (2016). Speech perception and generalization across talkers and accents. Linguistics: Oxford Research Encyclopedias.

Williams CE, & Stevens KN (1972). Emotions and speech: some acoustical correlates. J Acoust Soc Am, 52(4), 1238–1250. [PubMed: 4638039]

Witteman MJ, Bardhan NP, Weber A, & McQueen JM (2015). Automaticity and stability of adaptation to a foreign-accented speaker. Language and speech, 58(2), 168–189. [PubMed: 26677641]

Xie X, & Myers EB (2016). Sleep facilitates talker generalization of accent adaptation. Paper presented at the Annual Meeting of the Acoustical Society of America, Honolulu, Hawaii.

Xie X, Weatherholtz K, Bainton L, Rowe E, Burchill Z, Liu L, & Jaeger TF (in prep). Rapid adaptation to foreign-accented speech and its limits: A replication of Clarke and Garrett (2004).

Yu AJ, & Cohen JD (2008). Sequential effects: Superstition or rational behavior? Adv Neural Inf Process Syst, 21, 1873–1880. [PubMed: 26412953]

Zhang X, & Samuel AG (2014). Perceptual learning of speech under optimal and adverse conditions. Journal of Experimental Psychology: Human Perception and Performance, 40(1), 200. [PubMed: 23815478]

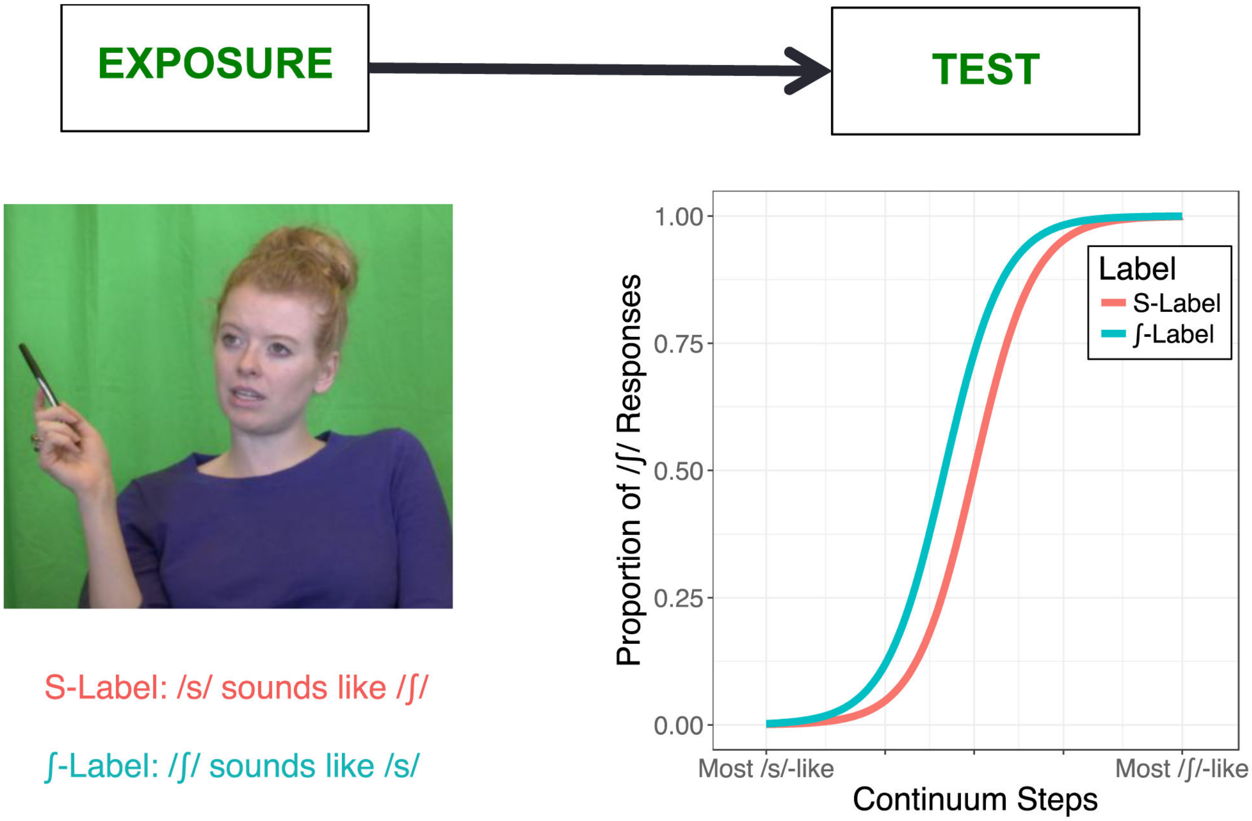

Figure 1:

Illustration of adaptation in a perceptual recalibration paradigm (also employed in the present work). Left side: During the exposure block, the S-Label and ʃ-Label group (between participants) hear shifted pronunciations from a single talker. Right side: Proportion of /ʃ/ responses as a function of continuum step in the S-Label and ʃ-Label conditions. The ʃ-Label condition is shifted towards /s/ (more /ʃ/ responses) and the S-Label condition is shifted towards /ʃ/ (more /s/ responses).

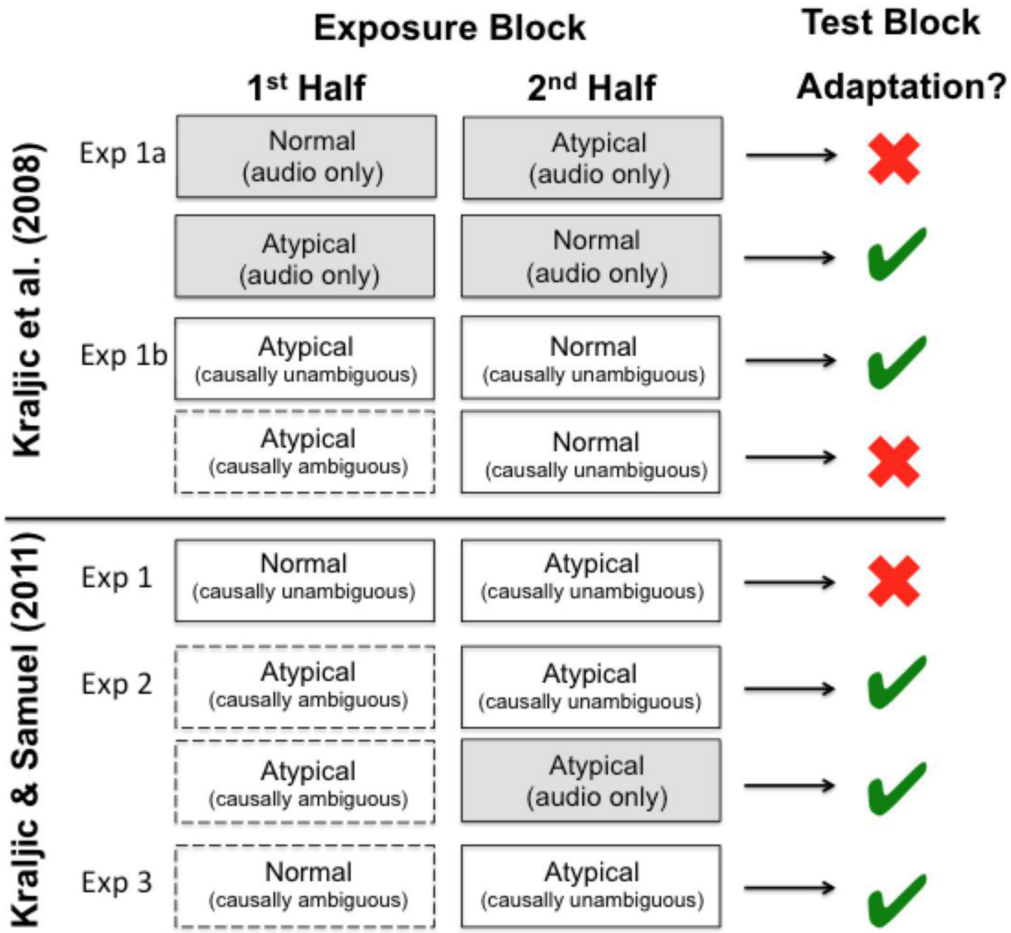
Exposure Block
Test Block
1st Half
2nd Half
Adaptation?
Kraljic et al. (2008)
Exp 1a
Normal (audio only)
Atypical (audio only)
✗
Atypical (audio only)
Normal (audio only)
✓
Exp 1b
Atypical (causally unambiguous)
Normal (causally unambiguous)
✓
Atypical (causally ambiguous)
Normal (causally unambiguous)
✗
Kraljic & Samuel (2011)
Exp 1
Normal (causally unambiguous)
Atypical (causally unambiguous)
✗
Exp 2
Atypical (causally ambiguous)
Atypical (causally unambiguous)
✓
Atypical (causally ambiguous)
Atypical (audio only)
✓
Exp 3
Normal (causally ambiguous)
Atypical (causally unambiguous)
✓

Figure 2:

Summary of conditions and results from Kraljic et al. (2008) and Kraljic & Samuel (2011). The rightmost column indicates the presence or absence of adaptation during the test block. All experiments manipulated what participants experienced in the first and second half of exposure (the block structure was opaque to participants). Exposure blocks either contained shifted atypical pronunciations (of either /s/ or /ʃ/, not shown) or only normal pronunciations. Presentation during exposure was audio only (shaded box) or audiovisual (non-shaded box). Critical pronunciations (either atypical or normal) were accompanied by pen-in-the-mouth videos (i.e., causally ambiguous, dashed border) or pen-in-the-hand videos (i.e., causally unambiguous, solid border).

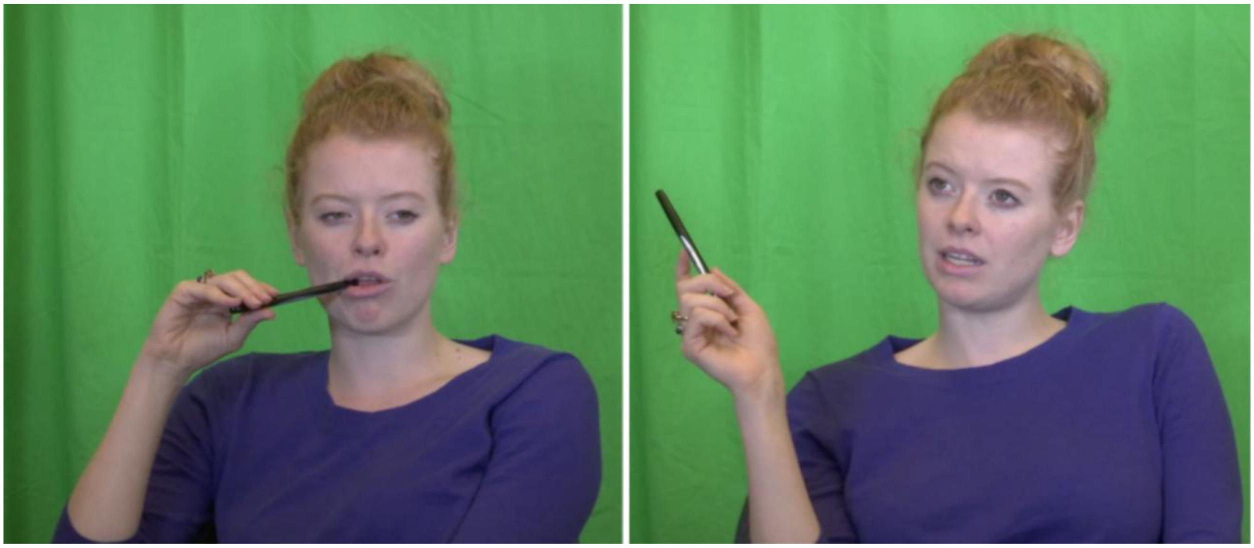

Figure 3:

Stills from example video stimulus. In this example, the speaker produces the shifted word 'initial' with a pen in her mouth (left) and a pen in her hand (right). The same audio file was used to dub each video in a pair. Reproduced with permission from Babel (2016).

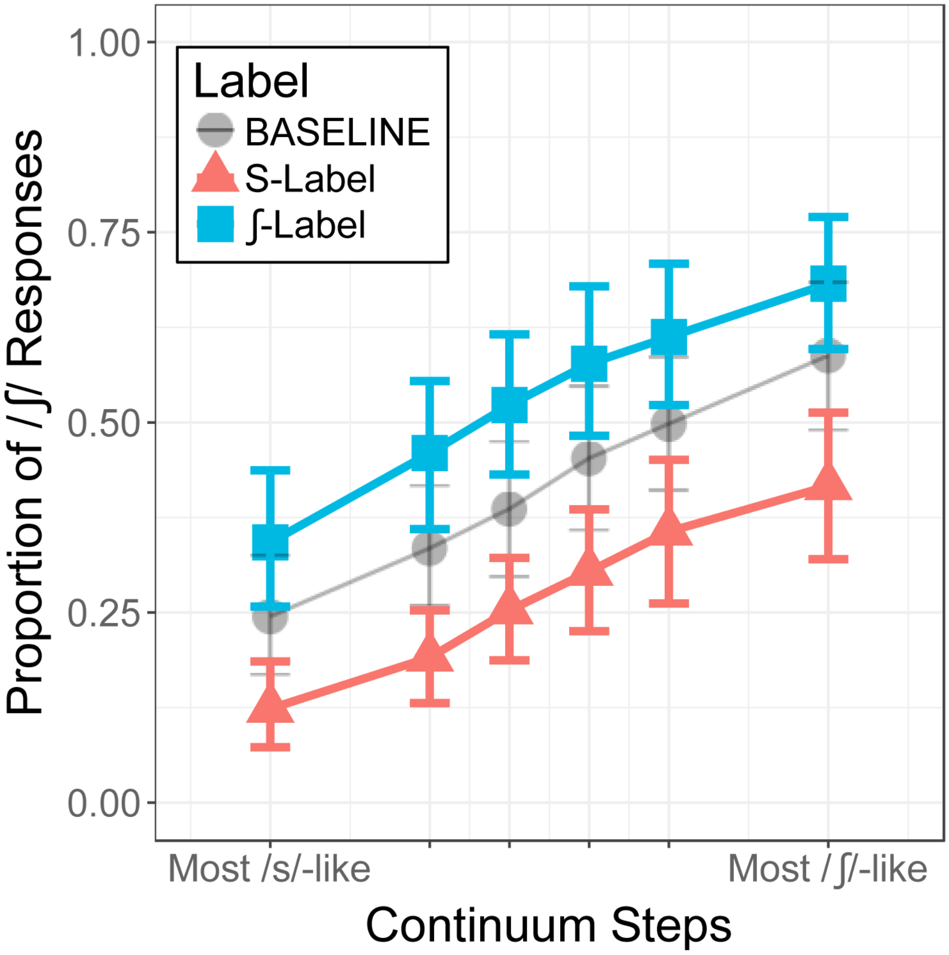

Figure 4:

Proportion of /ʃ/ responses as a function of continuum step in Experiment 1a. Participants in the ʃ - Label Condition (blue square) shift towards /s/ and participants in the S-Label Condition (red triangle) shift towards /ʃ/. The transparent black line represents the baseline condition, in which participants heard only normal pronunciations of all the critical tokens.

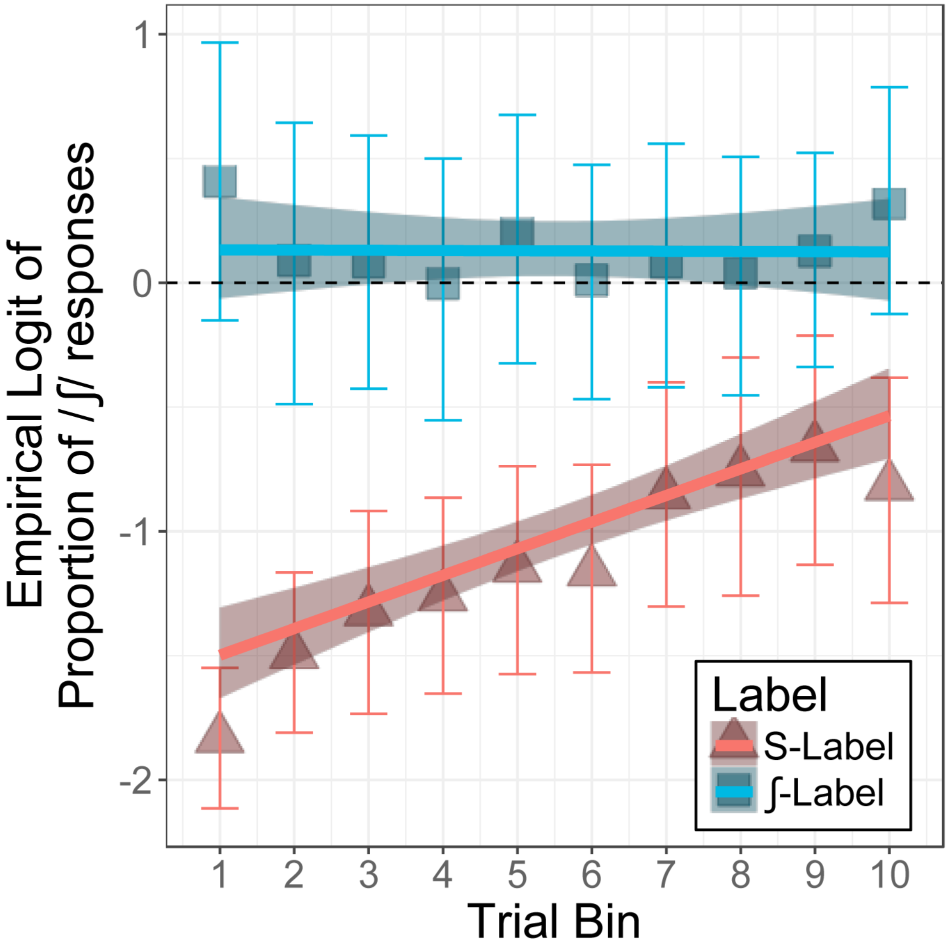

Figure 5:

/ʃ/ responses during test as a function of Trial Bin in Experiment 1a. Proportions of /ʃ/ responses were empirical logit transformed for ease of comparison with our analysis. Points show empirical data. Solid lines show predictions of the model we use to obtain corrected estimates of the category boundary shift at the *onset of test.* Both Label conditions move towards an empirical logit of 0 (dashed line) as a function of Trial Bin, reducing the difference between the two conditions in later trial bins.

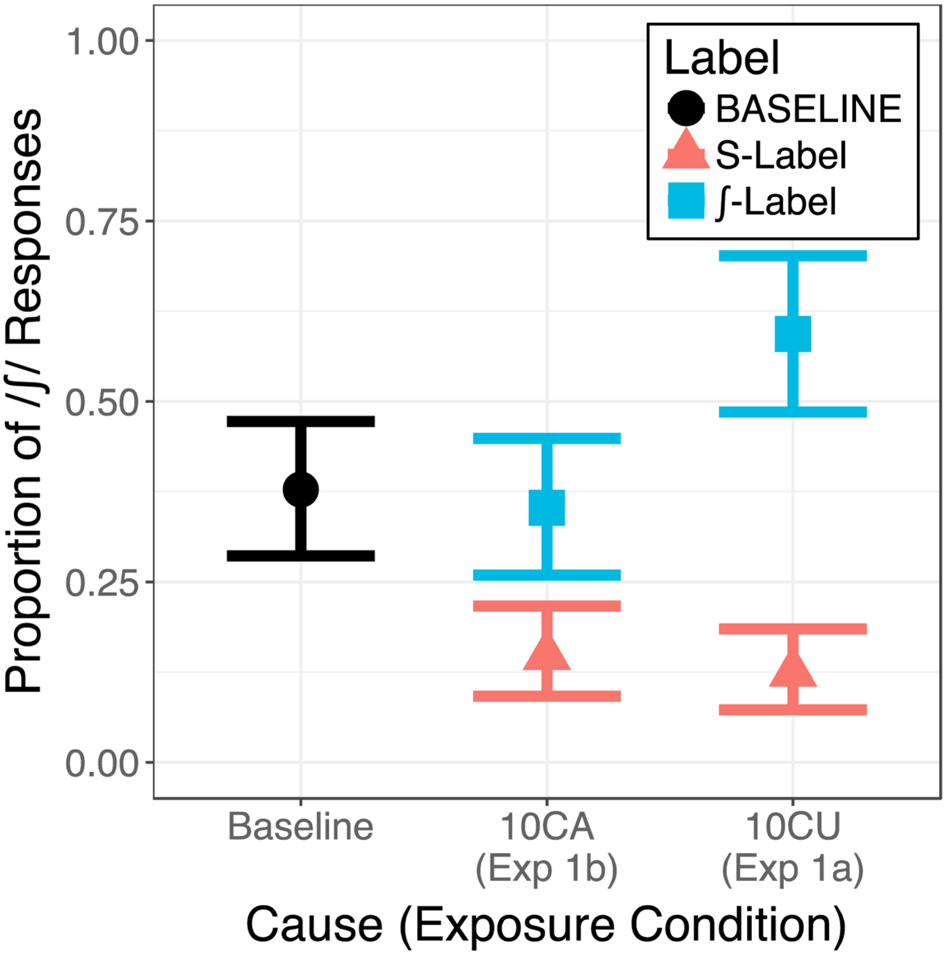

Figure 6:

Total /ʃ/ responses across conditions for the first trial bin and collapsing over /s/-/ʃ/ continuum steps in Experiments 1a and 1b. The baseline condition exposed participants to an equal number of /s/ and /ʃ/ as the other exposure conditions, except that both sounds were always pronounced normally. Comparison of the 10CU (10 Causally Unambiguous) condition to the baseline shows that participants adapt in both Label conditions following exposure to shifted stimuli paired with videos of the pen in hand. Comparison of the 10CA (10 Causally Ambiguous) condition to the baseline shows adaptation only for the S-Label condition; adaptation is blocked for the ʃ-Label condition (no difference compared to baseline).

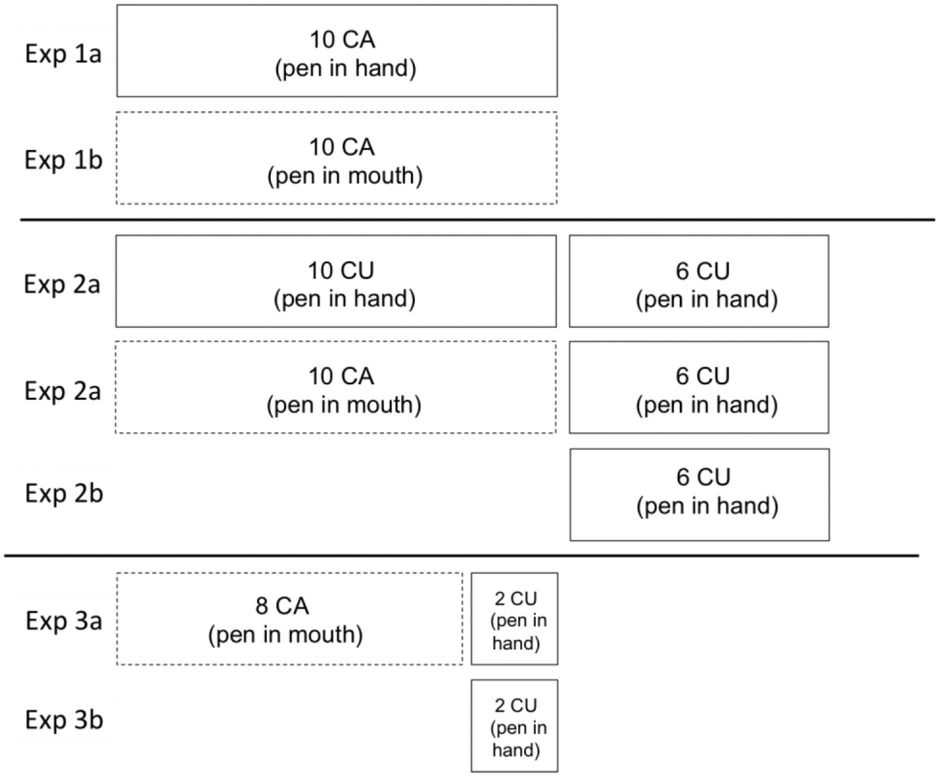

Figure 7:

A summary of the visual/causal evidence that accompanied the atypical critical items during the exposure block from our Experiments 1, 2, and 3. Percepts were either causally unambiguous (CU) or causally ambiguous (CA). The block structure was opaque to participants. The same filler-to-critical item ratio was maintained for all blocks (same for the word-to-non-word ratio).

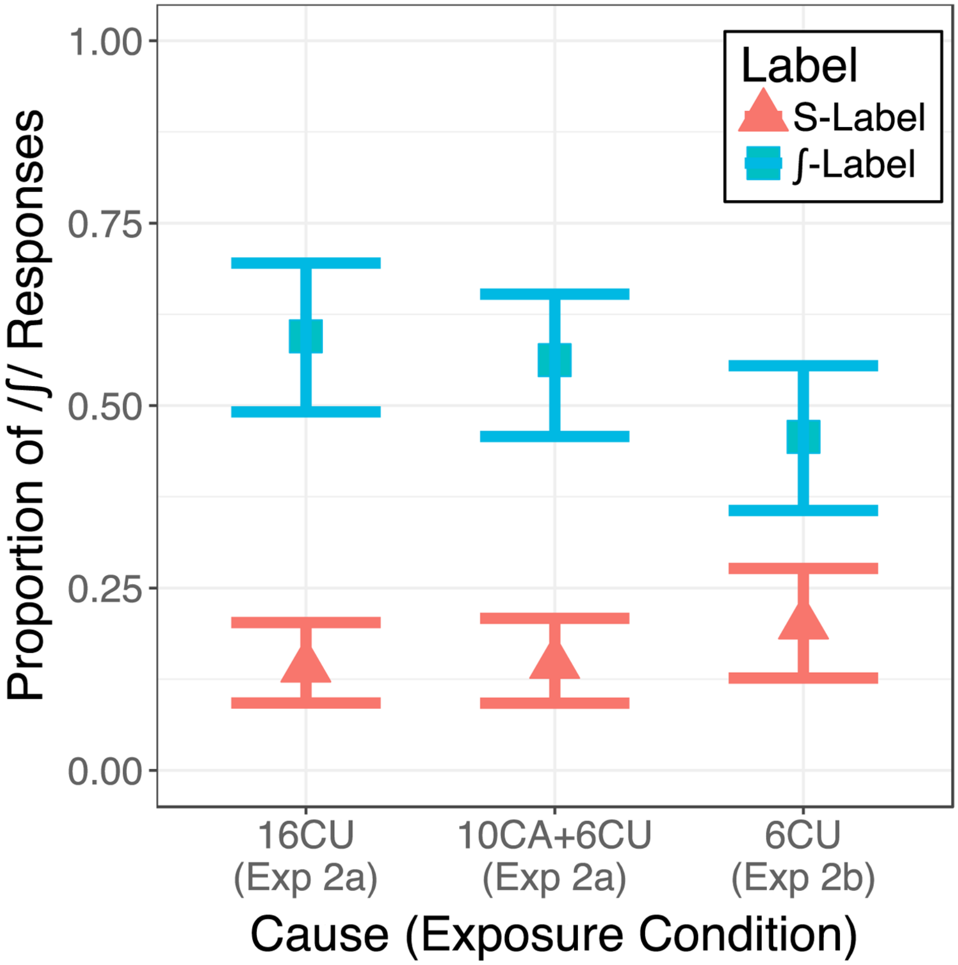

Figure 8:

Comparison of perceptual recalibration across exposure conditions of Experiments 1 and 2 at the first trial bin. Perceptual recalibration (difference between the S-Label and ʃ-Label groups) is observed in all conditions. However, the degree of perceptual recalibration is significantly smaller in the 6CU (6 Causally Unambiguous) condition than the 16CU (16 Causally Unambiguous) condition.

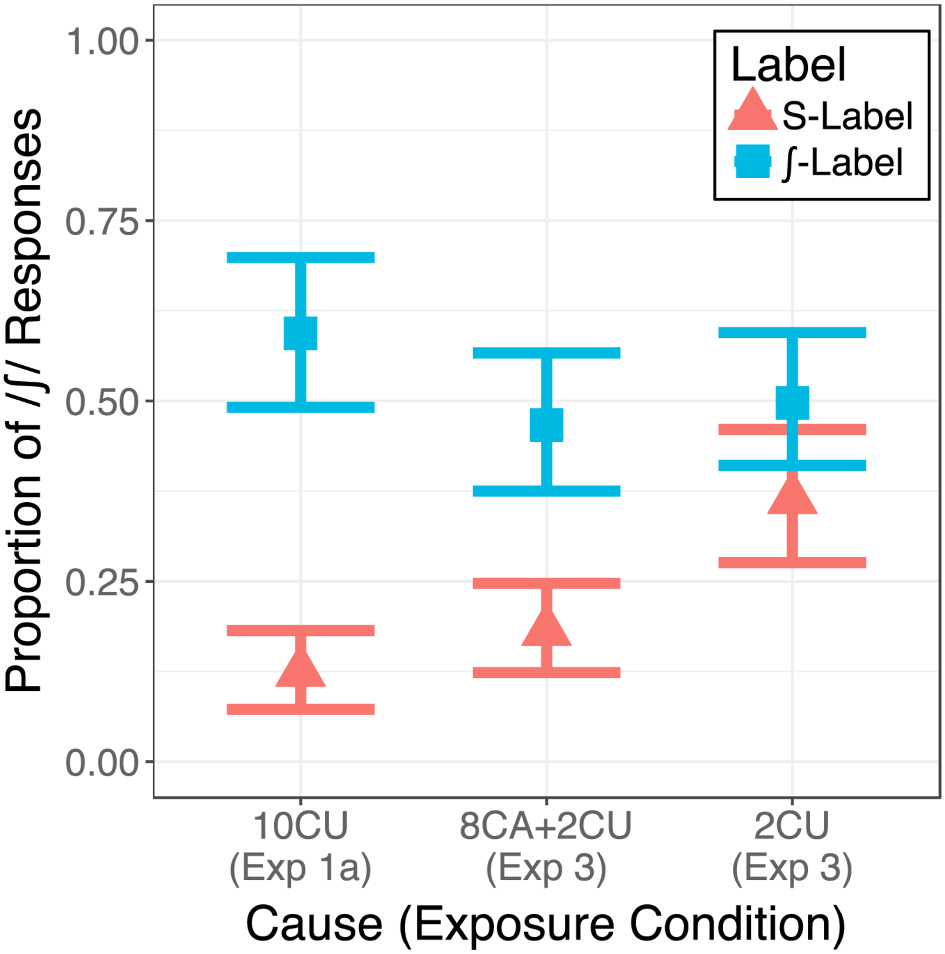

Figure 9:

Comparison of predicted perceptual recalibration effect at the first trial bin across Cause conditions of Experiments 1a and 3. Recalibration (difference between the S-Label and ʃ-Label groups) is observed in only the 10CU (10 Causally Unambiguous) and 8CA+2CU (8 Causally Ambiguous + 2 Causally Unambiguous) conditions.

Table 1

Breakdown of the number of excluded workers in all experiments. Each column represents one between-participant condition. We aimed for similar numbers of remaining participants per between-participant condition to rule out differences in power as explanation for differences in effects between those conditions. The comparatively high number of exclusions for Experiment 3b is due to the small number of exposure trials (making it more likely that 15% of those—i.e., 3 trials—are answered incorrectly).

Reason for exclusion	**Baseline**	**Exp 1a (10CU)**	**Exp 1b (10CA)**	**Exp 2a (16CU)**	**Exp 2a (10CA + 6CU)**	**Exp 2b (6CU)**	**Exp 3a (8CA + 2CU)**	**Exp 3b (2CU)**
Lexical decision accuracy	-	2	13	3	5	4	7	23
Repeat participant	3	1	3	4	-	4	-	-
No headphones	-	1	1	-	-	-	-	2
Catch question	-	2	1	1	-	1	-	3
Swapped keys	1	5	3	4	1	5	3	8
Multiple reasons	-	1	-	-	-	-	-	5
Total Exclusions	4 (6.3%)	12 (9.7%)	21 (15.9%)	12 (8.9%)	6 (5.0%)	14 (11.7%)	10 (7.8%)	41 (16%)
Remaining Participants	**60**	**112**	**112**	**123**	**115**	**108**	**118**	**113**

Table 2:

Experiment 1a - Mixed logit regression predicting proportion of /ʃ/ responses from Label (sum coded: ʃ-Label = 1 vs. S-Label = −1). The analysis included the maximal random effect structure justified by the design (by-participant intercepts). Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\hat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−0.97	0.26	−3.8	**<.001**
Label (ʃ vs. S)	1.14	0.26	4.4	**<.001**

Table 3:

Experiment 1a - Mixed logit regression predicting proportion of /ʃ/ responses from Label (sum coded: ʃ-Label = 1 vs. S-Label = −1), Trial Bin (with the first trial bin as 0), and their interaction. The analysis included the maximal random effect structure justified by the design (by-participant intercepts). Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\hat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−1.47	0.27	−5.5	**<.001**
Label (ʃ vs. S)	1.65	0.27	6.2	**<.001**
Trial Bin (First bin = 0)	0.10	0.01	8.5	**<.001**
Label:Trial Bin	−0.10	0.01	−8.8	**<.001**

Table 4:

Experiment 1a and 1b (combined)- Mixed logit regression predicting proportion of /ʃ/ responses from Cause (sum-coded: 10CU = 1 vs. 10CA = −1) Label (sum coded: ʃ-Label = 1 vs. S-Label = −1), Trial Bin (with the first Trial Bin as 0), and their interactions. The analysis included the maximal random effect structure justified by the design (by-participant intercepts). Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\hat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−1.84	0.18	−10.0	**<.001**
Cause (10CU vs. 10CA)	0.38	0.18	2.1	**<.05**
Label (ʃ vs. S)	1.28	0.18	7.0	**<.001**
Trial Bin (First bin = 0)	0.09	0.01	10.8	**<.001**
Cause:Label	0.36	0.18	2.0	**<.05**
Cause:Trial Bin	0.01	0.01	1.1	0.27
Label:Trial Bin	−0.07	0.01	−8.7	**<.001**
Cause:Label:Trial Bin	−0.03	0.01	−3.6	**<.001**

Table 5:

Experiment 1 – comparison to the baseline condition. Mixed logit regression predicting proportion of /ʃ/ responses from Exposure (treatment coded: baseline = reference level), Label (sum coded: ʃ-Label = 1 vs. S-Label = −1), Trial Bin (with the first Trial Bin as 0), and their interactions. In the baseline conditions, participants saw only causally unambiguous, normal pronunciations. The Exposure conditions consisted of the cross between ʃ-Label vs. S-Label with Causally Unambiguous vs. Causally Ambiguous, from Experiments 1a and 1b. The analysis included the maximal random effect structure justified by the design (by-participant intercepts). Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\widehat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−0.98	0.35	−2.8	**<.01**
ʃ-Label (10CA)	−0.33	0.52	−0.6	0.52
S-Label (10CA)	−2.17	0.51	−4.3	**<.001**
ʃ-Label (10CU)	1.16	0.50	2.3	**<.05**
S-Label (10CU)	−2.13	0.52	−4.1	**<.001**
Trial Bin (First bin = 0)	0.04	0.02	2.4	**<.05**
ʃ-Label (10CA):Trial Bin	0.00	0.02	0.1	0.91
S-Label (10CA):Trial Bin	0.09	0.02	3.7	**<.001**
ʃ-Label (10CU):Trial Bin	−0.04	0.02	−1.8	=*0.07*
S-Label (10CU):Trial Bin	0.17	0.02	7.2	**<.001**

Table 6:

Experiment 2a – Mixed logit regression predicting proportion of /ʃ/ responses from Label (sum coded: ʃ-Label = 1 vs. S-Label = −1), Cause (sum-coded: 16CU = 1 vs. 10CA+6CU = −1), Trial Bin (with the first Trial Bin coded as 0), and their interactions. (The analysis included the maximal random effect structure justified by the design (by-participant intercepts). Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\hat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−1.45	0.19	−7.8	**<.001**
Label (ʃ vs. S)	1.61	0.19	8.6	**<.001**
Cause (16CU vs. 10CA+6CU)	0.08	0.19	0.4	0.66
Trial Bin (First bin = 0)	0.03	0.01	3.1	**<.01**
Label:Cause	0.15	0.19	0.8	0.43
Label:Trial Bin	−0.08	0.01	−9.6	**<.001**
Cause:Trial Bin	−0.01	0.01	−1.6	=0.12
Label:Cause:Trial Bin	−0.00	0.01	−0.4	0.68

Table 7:

Experiment 2b: Mixed logit regression predicting proportion of /ʃ/ responses as a function of Cause (sum coded: 16CU = 1 vs. 6CU = −1 and 10CA+6CU = 1 vs. 6CU = −1), Label condition (sum-coded: ʃ-Label = 1 vs. S-Label = −1), Trial Bin (with the first Trial Bin coded as 0), and their interactions. Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\hat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−1.45	0.15	−9.7	**<.001**
Cause1 (16CU vs. 6 CU)	0.10	0.21	0.5	0.64
Cause2 (10CA+6CU vs. 6 CU)	−0.07	0.21	−0.3	0.74
Label (ʃ vs. S)	1.40	0.15	9.3	**<.001**
Trial Bin (First bin = 0)	0.02	0.01	2.9	**<.01**
Cause1:Label	0.34	0.21	1.6	*=0.10*
Cause2:Label	0.05	0.21	0.3	0.80
Cause1:Trial Bin	−0.01	0.01	−0.8	0.45
Cause2:Trial Bin	0.02	0.01	1.9	*=0.05*
Label:Trial Bin	−0.06	0.01	−8.8	**<.001**
Cause1:Label:Trial Bin	−0.02	0.01	−2.4	**<.05**
Cause2:Label:Trial Bin	−0.02	0.01	−1.7	*=0.09*

Table 8:

Mixed logit regression predicting proportion of /ʃ/ responses from Label (sum-coded: ʃ-Label = 1 vs. S-Label = −1), Cause (sum coded: 2CU = 1 vs. 8CA+2CU = −1 and 10CU = 1 vs. 8CA+2CU = −1), Trial Bin (with the first Trial Bin at 0), and their interactions. Rows that are critical to our analysis are highlighted in light gray for clarity.

Predictors	**Parameter Estimates**		**Significance Test**	
	Coef $\hat{\beta}$	**Std Err**	***z***	***p***
(Intercept)	−1.26	0.16	−8.0	**<.001**
Cause1 (2CU vs. 8CA+2CU)	0.53	0.22	2.3	**<.05**
Cause2 (10CU vs. 8CA+2CU)	−0.22	0.23	−1.0	0.32
Label (ʃ vs. S)	1.05	0.16	6.6	**<.001**
Trial Bin (First bin = 0)	0.07	0.01	9.6	**<.001**
Cause1:Label	−0.65	0.22	−2.9	**<.01**
Cause2:Label	0.62	0.23	2.7	**<.01**
Cause1:Trial Bin	−0.04	0.01	−3.8	**<.001**
Cause2:Trial Bin	0.04	0.01	3.8	**<.001**
Label:Trial Bin	−0.06	0.01	−8.8	**<.001**
Cause1:Label:Trial Bin	0.00	0.01	0.1	0.91
Cause2:Label:Trial Bin	−0.04	0.01	−4.6	**<.001**