**TO DO list:**

* Shawn:
  + Make exclusion table for Experiment 1a-c
* Gevher:
  + Update website and see whether it can be integrated into OSF (i.e., work locally on OSF, ideally with videos etc. being links to the local OSF repo)
* Florian:
  + Run analyses for experiment 1a-c with mo() (potential problem: interpretation of main effects, etc.)

Causal inference in audiovisual speech perception and adaptation

Shawn Cummings1,3, Gevher Karboga1, Menghan Yang1,4, and T. Florian Jaeger1,2

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

2Department of Computer Science, University of Rochester, USA

3Department of Speech, Language, and Hearing Sciences, University of Connecticut, USA

4Department of Psychological Sciences, University of Connecticut, USA

Word count (main text): ~XXX

Contact author:

Shawn Cummings

shawn.cummings@uconn.edu

University of Connecticut

D.C. Phillips Communication Sciences Building

Storrs, CT 06269

USA

**Author note:** This research was supported by a Bilski-Mayer Summer Research Fellowship to the first author and by the University of Rochester through funding for *BCS 206: Undergraduate Research in Cognitive Science*. In this yearlong class, undergraduate students conduct their own replications of important findings in the cognitive sciences, in collaboration with faculty mentors.

Earlier versions of this work were presented at the University of Rochester Undergraduate Research Expo, and as a poster at the 62nd annual meeting of the Psychonomic Society. We owe many thanks to other students and instructors of BCS 206, who provided invaluable feedback throughout the project. We gratefully acknowledge that this project is made possible through several researchers’ dedication to open science: the audio recordings were obtained from Tanya Kraljic and Arthur Samuel, the video recordings were obtained from Molly Babel; the Javascript code for the experiment is based on openly accessible code originally developed by Dave Kleinschmidt, with input from Linda Liu and Zach Burchill (Human Language Processing Lab, University of Rochester, URL). Data collection via Prolific was facilitated by Proliferate, a tool developed and maintained by Sebastian Schuster (ALPs Lab, Stanford).

**Author contributions:** SNC proposed the experiment. SNC, GK, and MY designed the experiment and developed the hypotheses, with input from TFJ. SNC created the audiovisual stimuli from the source audio and video files. TFJ and GK programmed the web-based experiments. GC created a webpage with links to all experiments for demonstration purposes. MY and TFJ conducted data visualization and organization, with input from SNC. TFJ conducted the statistical analyses. All authors jointly interpreted the results. SNC and TFJ wrote the paper, with input from all authors.

**Abstract**

XYZ

**Keywords**: speech perception; perceptual recalibration; causal inference; compensation for coarticulation; XXX

**Open science statement:** The documented Javascript code, experimental lists, and audiovisual materials for all experiments are available via OSF at XYZ. The same holds for all analysis code and trial-level data, which are shared in the form of “knittable” R Markdown (REF).Live versions of all experiments can be called at XYZ.

# Introduction

Understanding one another is paramount to daily life. And yet, the mechanisms underlying our ability to do so are not yet fully understood. Speech involves a ‘paradox of perception’ (Kraljic, Samuel, & Brennan, 2008), in that listeners must be able to discern miniscule acoustic differences (e.g., the deviations in frequency spectra between English “f”, “th”, “s”, and “sh”) as meaningful, but simultaneously must perceive potentially far greater acoustic differences as non-contrastive. Indeed, one talker’s production of, for example, the “s” sound in *sun* may well be acoustically identical to a different talker’s “sh” in *shun* (Newman et al., 2001, hereafter, /s/ and /ʃ/, respectively). Similarly, a single talker may generate the same acoustics for an /s/ in one phonetic context as for an /ʃ/ in another (Smits, 2001). Both of these factors contribute to the ‘lack of invariance’ problem, which has haunted research in speech perception for decades and continues to be a central theme of current work (Liberman et al., 1957; Saltzman et al. 2021). While there are certainly more factors confounding a one-to-one mapping between acoustics and meaning (noise and rote chance being obvious examples), the bulk of investigation in speech perception has focused on mechanisms by which humans overcome either contextual variation within one talker’s speech, or rather differences in speech between talkers. These lines of research have been pursued largely independent from one another, and their effects have been attributed to largely separate mechanisms.

The current investigation is situated within the realm of differences in speech between talkers, and follows a large extant literature highlighting adaptation as a mechanism by which the lack of invariance is ameliorated. Speech from a novel talker, especially whose realizations of phonetic categories don’t align with a listener’s prior experience, is more difficult and time consuming to comprehend (e.g. Clarke & Garrett, 2004; Bradlow & Bent, 2008). Listeners, however, are able to rapidly adjust their expectations to better reflect the input they receive. This is true for differences in production resultant of real-world talker characteristics, such as physiology, sociophonetic traits, and idiosyncratic pronunciation habits, as well as for experimenter manipulation of specific sound categories (e.g. Norris, McQueen, & Cutler, 2003). Anecdotally, however, one can imagine that adjusting expectations towards a talker’s speech is *only* facilitatory to future comprehension if the perceived variation is likely to re-occur. While *characteristic* talker differences should elicit adaptation, the overarching goal of efficient communication is not served via adaptation to transient or incidental variations. A growing body of literature, which the current investigation hopes to contribute towards, has examined the specificity of adaptation with regard to *characteristic* versus *incidental* talker variation (Kraljic, Samuel, & Brennan, 2008; Kraljic & Samuel, 2011; Liu & Jaeger, 2018; Liu & Jaeger, 2019; White & Daub, 2021). Specifically, articulatory obstructors, such as a lollipop or pen in a talker’s mouth, have been found to reduce or block talker-specific learning, posited as potentially resultant of causally attributing the sound shifts to the object rather than to the talker (White & Daub, 2021; Kraljic, Samuel, & Brennan, 2008).

The situation created by studies on adaptation in the context of incidental causes provides a productive but complicated test case. Listeners are simultaneously provided with two variables. On the one hand, they hear a novel talker exhibiting speech deviant from their expectations. This has been found in upwards of 80 manuscripts to elicit adaptation (e.g. Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2005; Eisner & McQueen, 2006; van Linden & Vroomen, 2007; Liu & Jaeger, 2018; Cummings & Theodore, 2022; Cummings & Theodore, 2023). On the other, they hear speech within a specific abnormal context. To date, studies have not examined listeners’ ability to overcome this contextual variation in the abstract, outside of talker-specific learning. It is well established, however, that listeners *do* account for contextual variation within one talker’s speech**. The current investigation hopes to further understanding of the specificity of learning through acknowledging the potential joint contributions of both talker and context on variability in speech.** To this end, an initial goal is to establish incidental causality as a type of contextual variation that perceptual mechanisms may already be equipped to handle. This is achieved through a brief review of findings on compensation for contextual variation, followed by a perception experiment without a learning component (Experiment 1). Having established this link, Experiments 2 and 3 then extend prior study *within* learning. Extant findings herein, as well as specific theoretical implications and proposals, will be unpacked after Experiment 1.

# Experiments 1a-c

The phonetic realization of sound categories is affected by their phonetic context, a process known as co-articulation. For example, English fricatives have a lower spectral center of gravity directly following the vowel /u/ (as in *moon*) compared to the vowel /XXX/ (Soli, 1981; Yeni-Komshian & Soli, 1981). As the spectral center for /ʃ/ is generally lower than that of /s/ in English (Jongman, Wayland, & Wong, 2000), the presence of /u/ serves to make /s/ segments acoustically more similar to typical /ʃ/ segments. Human speech perception is known to compensate for this: for a fricative ambiguous between /s/ and /ʃ/, the presence of a preceding /u/ biases listeners towards /s/ responses (Mann & Repp, 1980; Mann & Soli, 1991). That is, listeners seem to attribute lowered spectral center of gravity at least in part to the co-articulatory effect of the preceding /u/, rather than an intention to produce a /ʃ/ (Fowler, 2006). This relation between co-articulation in production and compensation in perception is schematically illustrated in Figure XXX. Similar compensation effects have now been documented for a wide range of phonetic contexts (for review, see REF), often under the alternative term normalization (e.g., REFS).

There is also evidence that compensation for co-articulation is not limited to acoustic context. For example, in an effect closely resembling that of preceding /u/, visual presence of lip-rounding preceding audio of an ambiguous /d-g/ blend biases listeners towards perceiving /g/ (Fowler, et al., 2000; Holt et al., 2005). Lip-rounding lowers the third formant (F3). In the absence of this visual context, lower F3 would be more likely to result from producing /d/ rather than /g/. Paralleling compensation for preceding /u/, listeners thus seem to compensate for the visual context of lip-rounding (see also Viswanathan & Stephens, 2016). Results like these have led to the hypothesis that compensation can occur regardless of the type and modality of contextual cues. As Fowler (2006, p. XXX) put is: compensation for lip-rounding would be equally expected if a talker “was about to whistle a merry tune or about to kiss a loved one”, as “it does not matter why the lips were rounded; it only matters that they were rounded [prior to articulation of the /d/-/g/ sound] and, therefore, would lower the F3 of the syllable that the gesture overlapped with temporally” (Fowler, 2006).

In line with this hypothesis, visual evidence of articulatory obstruction has also been found to affect the perception of the phonetic contrast we focus on here: the presence of a bite block has been found both to impact the acoustics of /s/ and /ʃ/ productions, and to elicit compensation in perception (McFarland & Baum, 1995; Baum et al. 1996). The presence of a pen in the mouth could well serve a similar function, but has not to date been examined to date.

A pen in the mouth has two visually apparent effects on the articulators during speech. The first of these is to increase the opening of the jaw and size of oral cavity (as the pen prevents the mouth from closing), and the second is to force lip rounding around the protruding end of the pen. As the size of the oral cavity opening and amount of air constriction are inversely related for fricatives, forced mouth opening is expected to lower spectral center of gravity. Lip rounding is also expected (in English) to lower the spectral center of gravity for surrounding fricatives. As lower spectral center of gravity is one of the primary cues distinguishing /ʃ/ from /s/ in English (REF), both of these effects are predicted to make fricatives produced with a pen in the mouth acoustically more ‘/ʃ/-like’. If listeners compensate for either or both of these effects of the pen on articulation, it should bias their perception towards /s/ (against /ʃ/).

Experiments 1a-c test this prediction. All three sub-experiments employ the exact same design and procedure but differ in the specific visual and acoustic stimuli they employ, as well as minor details of the post-experiment survey. Specific differences are described in the methods. Participants were presented audiovisual speech stimuli which formed six steps along a continuum from *ashi* to *asi.* Audio was dubbed onto video of a young female talker holding a pen (see Figure 1). During the production of the critical /s/-/ʃ/ fricative, the talker either had the pen in her mouth (Figure 1, left) or rather in her hand outside of the mouth (Figure 1, right). We were interested in whether the presence of the pen—or its visually apparent effects on the articulation of /s/ and /ʃ/—affects the interpretation of acoustic cues to the /s/-/ʃ/ contrast. Participants performed an identification (categorization) task, answering whether they thought the talker in the video produced *ashi* or *asi*.



Figure 1 Stills of video components of two audio-visual items in Experiments 1a-c, illustrating the critical manipulation. Participants saw and heard audio-visually presented speech stimuli drawn from an acoustic “asi” to “ashi” continuum. During the production of the fricative, the talker either had the pen in the mouth (left) or in the hand (right).

The use of audiovisual stimuli comes with unique challenges. While our goal was to investigate how the presence of the pen affects the perception of the acoustic input, the use of audiovisual stimuli entails that participants also had access to visual cues to the /s/-/ʃ/ contrast, such as lip-rounding (Proctor, Shadle, & Iskarous, 2006). Speech perception is well-known to integrate acoustic and visual information to articulation, and identification responses are known to reflect this integration (McGurk & McDonald, 1976; see also Bejjanki et al., 2011; Franken et al., 2017; Lüttke et al., 2018). This raises questions about how the presence of visual cues to the articulation of /s/ or /ʃ/ in the video affect participants’ identification responses. One way to address this question would be to manipulate the video stimuli—either by holding them constant or by gradiently varying the visual cues to /s/ and /ʃ/, independent of the auditory cues. We decided against the second possibility primarily for reasons of feasibility. It is substantially more difficult to create ecologically valid visual manipulations than it is to create ecologically valid acoustic manipulations. The few previous studies that *have* gradiently manipulated visual cues to articulation have either employed animation, rather than real-life videos, or a single ‘ambiguous’ real-life video (e.g. Bejjanki et al., 2011; Kang, Johnson, & Finley, 2016). None of the studies that we are aware of have manipulated video while also modeling the visual consequences of perturbing articulation by a pen is in the mouth.

We thus decided to take an alternative approach. We created the video stimuli by extracting short segments from video recordings of the talker pronouncing words that contained *asi* or *ashi*-like sequences (e.g., *democracy*, which ends in a sound sequence highly similar to *asi*). This means that the audiovisual stimuli in Experiments 1a-c contain visual information that is expected to affect participants’ identification responses. For the test item derived from an original video recording of *democracy,* for example, we would expect responses to be biased towards *asi*. For a video extracted from a video recording of *machinery*, on the other hand, we would expect responses to be biased towards *ashi*. The design of Experiments 1a-c therefore fully crossed the visual /s/ or /ʃ/ bias of the original video clip with the synthesized acoustic *ashi–asi* continuum and the location of the pen. This resulted in a 2 (visual /s/- vs. /ʃ/-bias) x 6 (steps along acoustic /s/-/ʃ/ continuum) x 2 (pen in mouth vs. hand) design, with all conditions being manipulated within participants.

## Methods

Except for the use of audiovisual rather than audio-only stimuli and minor procedural changes reported below, Experiments 1a-c closely followed the norming experiments in Liu & Jaeger (2018).

***Participants.*** Following Liu and Jaeger (2018), participants were recruited from Amazon's crowdsourcing platform Mechanical Turk. Each experiment recruited 64 participants, balanced across two lists that balanced nuisance factors and are described below.

Participants were paid $6/hour prorated by the duration of the experiment (15 minutes). Participations only saw the experiment advertised, and could only participate in it, if (i) they were located within the US, (ii) had an approval rating of 99% or higher, and (iii) had not previously taken any of the other experiments in this study or in Liu and Jaeger (2018, 2019). Additionally, participants who did not fulfill the software requirements (a recent version of the Chrome browser engine) could see the recruitment, but could not participate in the experiment. Participants allowed to complete the experiment but were excluded from the analysis if they met any exclusion criteria, described below.

***Materials.*** To create the audiovisual stimuli, we combined audio and video recordings.

*Audio recordings.* The acoustic stimuli for all three experiments were selected from the same 31-step acoustic continuum from *ashi* to *asi* created by, and used in, Liu and Jaeger (2018). This continuum was created with FricativeMakerPro (McMurray, Rhone, & Galle, 2012) based on recordings of typical *ashi* and *asi* pronunciations by a female talker in her twenties—the same recordings elicited and employed in Kraljic et al. (2008). The audio recordings for the present study thus come from the same female talker used in previous research on perceptual recalibration (Kraljic & Samuel, 2005; Kraljic et al., 2008; Kraljic & Samuel, 2011, Liu & Jaeger, 2018). As in Liu & Jaeger (2018), we selected six steps along the 31-step continuum. These steps were selected with the goal to maximize the ability to detect effects of both the acoustic continuum and our other manipulations. The latter includes the location of the pen, but also the perceptual recalibration effects that we investigate in Experiments 2 and 3: as mentioned above, the secondary purpose of Experiments 1a-c was to norm the test phase for those perceptual recalibration experiments. To detect effects of the acoustic continuum, it is important for the test locations to span a sufficiently large range along the continuum. However, the power to detect other effects—including the location of the pen—is highest at test steps that elicit close to 50% *ashi* and 50% *asi* responses. Similar to experiments on perceptual recalibration, we thus aimed to select one step that, across all other manipulations, would yield approximated 25% *asi* responses, four steps that would yield close to 50% *asi*-responses, and one step that would yield 75% *asi* responses (e.g., Kraljic et al., 2008 and subsequent work).

The six acoustic continuum steps selected differed between Experiments 1a-c. The intention for Experiment 1a was to employ the exact same acoustic steps as in Liu and Jaeger (2018): 13, 15, 16, 17, 18, and 20, where higher numbers indicate acoustically more “ashi”-like steps. However, a labelling mistake resulted in the steps being internally named with higher numbers indicating proportion of asi, rather than “ashi”. The actual steps that were used in Experiment 1a were 12, 14, 15, 16, 17, and 19, unintentionally shifting the test continuum 1 step towards the “asi” end, compared to Liu and Jaeger (2018). Indeed, as we report below, Experiment 1a yielded overall more asi than “ashi” responses. Experiment 1b thus attempted to expand the test stimuli towards the “ashi” end of the continuum. However, as we still had not discovered the labeling mistake, we ended up expanding the continuum even further towards the asi end instead, with steps 10, 13, 14, 15, 16, 20. At this point, we identified the labeling mistake. Experiment 1c employed steps XYZ [compared to LJ18]. We emphasize that the location of the test steps is *not* critical for Experiments 1a-c.

*Video recordings.* The videos for the test stimuli were extracted from the exposure videos employed in the perceptual recalibration experiments in Liu and Jaeger (2018). These videos were recorded for Babel (2016) and generously provided by Dr. Molly Babel, as the original video stimuli from Kraljic et al. (2008; Kraljic & Samuel, 2011) are no longer available. The videos show a female talker of similar age as the one employed in the audio and video recordings of Kraljic et al. (2008; Kraljic & Samuel, 2011), providing a highly plausible match for the voice of the talker in audio recordings. Previous study has used this exact combination of the videos recorded by Babel and the audio recorded by Kraljic and found perceptual recalibration for audio-only test stimuli from the audio recorded by Kraljic (Liu & Jaeger, 2018).

The stimuli created by Babel and colleagues (Babel 2016) did not contain video recordings of the *ashi-asi* nonce-words, and the talker recorded by Babel and colleagues was no longer available (Molly Babel, p.c. July 17, 2020). For Experiment 1a, we thus identified twelve video recordings of exposure stimuli with the required sound sequence similar to *ashi* (e.g. *m[achi]nery*) or *asi* (e.g. *democr[acy]*) (for full list, see supplemental materials). Care was also taken to select recordings for which the relevant sequence matches the duration of the acoustic stimulus. We used the open-source video editing software Shotcut (shotcut.org) to extract the relevant video sequence out of the original recordings. Following the procedure used in previous work to create the exposure videos (Babel 2016), we added a fade-in and fade-out (each of 300 msecs) to the beginning and end of the new video segments. This resulted in videos of, on average, 1361 msecs duration (SD = 54 msecs).

Half of the twelve videos were extracted from video recordings of the talker pronouncing a word with an *asi* sequence (e.g., *leg[acy]*, henceforth visual /s/-bias). The other half were extracted from video recordings of the talker pronouncing a word with an *ashi* sequence (e.g., *gl[aci]er*, henceforth visual /ʃ/-bias). For each of those six videos, half showed the talker with the pen in the mouth and half showed the talker with a pen in the hand. The presence of a visual bias towards /s/ or /ʃ/ and the location of the pen were thus fully crossed between the twelve video items.

Experiments 1b and 1c employ eleven of these twelve videos. The twelfth video was replaced with a video in kind because, as described below, the results of Experiment 1a indicated a particularly strong visual bias for that video.

*Audio-visual stimuli.* The audio and video recordings were combined into audiovisual stimuli following the same procedure used in Liu and Jaeger (2018). Care was taken to ensure that the audio and video recordings aligned. We fully crossed the six steps along the acoustic continuum with each of the 12 video items, resulting in 72 audiovisual stimuli for each of the three experiments. The original video and audio recordings as well as all audiovisual test stimuli for both experiments are available as part of the OSF repository.

***Procedure.*** Our web-based procedure closely followed Liu and Jaeger (2018). The experiment consisted of (1) instructions, followed by (2) a test phase and (3) a post-experiment survey. Exact replica of all experiments for demonstration purposes can be found at XYZ. The same page contains screen recordings of a participant taking the experiment.

*Instructions.* The first page of instructions informed participants “This HIT is a psychology experiment about how people understand speech. Your task will be to listen to words, and to press a button on the keyboard to tell us what you heard.” Participants were informed that "It is extremely important that you use over-the-ear headphones of good sound quality for this experiment. If your headphones cost less than $30, then it is likely that they do not fulfill our criteria.” Participants were informed of the duration of the experiment, payment, eligibility, then completed a sound check, and gave consent. Following all previous experiments in our lab, these steps were all available prior to accepting the HIT, but in order to start the experiment, participants had to accept the HIT.

*Test phase.* At the beginning of the test phase, participants were instructed:

*You will see and hear videos of a female speaker producing words. Your task is to decide whether the speaker is saying “asi” or “ashi”. We appreciate your attention to this task. Please answer as quickly and accurately as possible, without rushing. You may hear similar sounds several times. As a form of quality control, you may sometimes see a white dot in the video. If it occurs, it is easy to see. If you see a white dot, please press “B” instead of answering. Do not press “B” unless you see a white dot. This helps us distinguish you from a robot.*

The instructions about the catch trial were included for the sake of comparability with Experiments XXX-XXX. None of the trials during the test phase actually contained a white dot. Participants then completed 72 trials of an 2AFC identification task. Participants could respond *asi* or *ashi* (via the X and M keys on their keyboard) only after the video had finished playing. Catch trial responses could be registered at any point during the video and caused the video to stop and the next trial to start. A progress bar indicated how many trials had been completed and how many remained, and the key binding was indicated at the top of the screen (see Figure 3). Key binding was counterbalanced across participants. This was the only nuisance variable, resulting in two between-participant lists. Each trial ended by the participant pressing M, X, or B (to indicate a catch trial). Both the response and the response time were recorded.

A child holding a pen

Description automatically generated with low confidence

Figure 2. Screenshot of 2AFC identification trial during Experiments 1a-c, identical to test phase of Experiments 2 and 3. The progress bar (experiment was roughly 30% completed at time of screenshot) and key bindings were visible throughout the experiment.

The order of test stimuli was determined separately for each participant through constrained randomization that grouped stimuli into blocks and then randomized the order within and across blocks (Kraljic et al., 2008; Liu & Jaeger, 2018). Specifically, the 72 audiovisual stimuli were grouped into six blocks of 12 stimuli so that each of the 12 video items occurred exactly once within each block. Each block further fully crossed the two pen locations (pen in hand vs. mouth) with the two visual bias conditions (/s/ vs. /ʃ/), resulting in 3 video items each for each of these four conditions. Each block of 12 stimuli further consisted of two instances each of each of the six audio conditions (steps along the asi – “ashi” continuum). One of these two instances occurred with the pen in the mouth and one occurred with the pen in the hand. Across the six blocks all 72 combinations of the 12 video items and the six audio conditions occurred exactly once. The order of the 12 test stimuli within each block was fully random.

*Post-experiment survey.* The post-experiment survey for each experiment is available as part of the OSF repository. The post-experiment survey for Experiment 1a was identical to that of Liu and Jaeger (2018). Questions assessed the quality of the audio equipment and whether participants experienced stalling of audio or video (to help us catch problems without code). Of particular note, the survey also contained a catch question, asking about the gender of the talker shown during the test phase. Following this survey, a final exit survey collected demographic information using the gender, age, race, and ethnicity categories required for NIH reporting. All responses in the demographic survey were indicated as optional.

For Experiment 1b we changed the post-experiment survey. We removed one question about audio quality which asked participants to rate the quality of their audio equipment as “good, professional, bad etc.”. Previous analyses in our lab have found that the subjective nature of this question makes it hard to interpret. And, in order to encourage the participants to answer survey questions truthfully, we emphasized that they would get compensated independent of their answers, and that truthful answers would greatly help us with the interpretation of results. For Experiment 1c, we made a few additional changes. [describe changes here.]

***Exclusions.*** We followed all applicable exclusion criteria of Liu and Jaeger (2018). Following Liu and Jaeger (2018), we removed (1) participants who experienced technical difficulties and did not complete the experiment and (2) participants who did not answer the catch question about the talker's gender correctly. We also excluded participants who likely had swapped the response keys, as determined by their identification function. For this purpose, we considered participants with significant slopes in the opposite of the expected direction as likely having swapped the response keys.[[1]](#footnote-2) No other criteria from Liu and Jaeger (2018) applied to Experiments 1a-c. Table 1 summarized the number of participants excluded by each criterion, and the participants remaining for analysis. Since the present experiments are likely high-powered, compared to previous replications, and since the analysis approach we employ below does not require balanced data, we decided not to replace excluded participants with newly recruited participants.

[INSERT TABLE]

Table 1 Summary of participant exclusions for Experiments 1a-1c.

## Results

Experiments 1a-c served two purposes, the first being theoretical—to test whether listeners compensate for the (visually evident) effects on articulation of a pen in the mouth—and the second methodological—to identify audiovisual test tokens suitable for the remaining experiments, i.e., stimuli that are centered around 50% *ashi* responses, ranging from about 25-75% *ashi* responses?

***Analysis approach.*** Here and throughout this study, we use Bayesian generalized linear mixed models with a Bernoulli (logit) link—mixed-effects logistic regression—for the analysis of identification responses (for an introduction to mixed-effects logistic regression, see Jaeger, 2008). In addition to the three design variables (visual bias, acoustic continuum, and pen location), we followed Liu and Jaeger (2018) and included test block as a predictor in the analyses (ranging from 1-6). This was done because previous work has found that responses can change throughout the test phase. Specifically, responses have been found to trend towards a uniform distribution of 50% *ashi* and 50% *asi* responses during continued testing (Liu & Jaeger, 2018, 2019; Tzeng, Nygaard, & Theodore, 2021).

Responses (1 = “ashi” vs. 0 = asi) were regressed against visual bias (effect coded: .5 = /ʃ/-bias vs. -.5 = /s/-bias), acoustic continuum (continuous, centered), pen location (effect coded: .5 = in mouth vs. -.5 = in hand), and test block (continuous, with block 1 coded as 0) as well as all their interactions. By coding block 1 as 0 (and block 2 as 1, etc.), the effects of all other variables are assessed in the first block of the experiment.

Following recommendations by Gelman (2008), the two continuous predictors were standardized by diving through twice the standard deviation. This puts the continuous predictors on a similar scale as the categorical predictors, and makes sure that the regularizing priors affect the estimation of all effects in similar ways. The analyses further contained the full random effect structure (by-participant intercepts and slopes for all predictors; by-video item intercepts and slopes for acoustic continuum since visual bias and acoustic continuum were manipulated between items).

All analyses were fit using the library *brms* (Bürkner, 2017) in R version 4.1 (REF). We followed recommended practice and use weakly regularizing priors to facilitate model convergence, the exact same as in our previous work (e.g., Hörberg & Jaeger, 2021; Xie, Liu, & Jaeger, 2021). For fixed effect parameters, we use Student priors centered around zero with a scale of 2.5 units (following Gelman et al., 2008) and 3 degrees of freedom. For random effect standard deviations, we use a Cauchy prior with location 0 and scale 2, and for random effect correlations, we use an uninformative LKJ-Correlation prior with its only parameter set to 1 (Lewandowski et al., 2009), describing a uniform prior over correlation matrices. Model diagnostic indicated convergence (e.g., all ).

Chart, box and whisker chart

Description automatically generated

Figure 4. Average “ashi” responses for each of 12 video frames in Experiments 1a-c (panels A-C, respectively). Results are collapsed across the entire acoustic continuum. Frame 5 was replaced by Frame 13 after Experiment 1a. As intended, Frame 13 is numerically less biasing.

Experiments 1a-c additionally investigated the biasing effect of visual artifacts of the original sound produced in video test tokens. These effects can be seen in Figure 4 (Panel A). Unsurprisingly, visual inspection reveals that videos originally of a talker producing /ʃ/ sounds yielded more /ʃ/ responses. Within the /ʃ/-label video frames, Frame 5 was especially biasing. This token was therefore re-adjusted in Shotcut between Experiments 1a and 1b. Figure 4 (Panels B & C) shows the biasing effects in Experiments 1b-c, including the new video frame.

Finally, we were interested in how effects of pen might be limited to certain visually specified phonetic contexts. Figure 5.

Lip rounding seems to effect /ʃ/ more than /s/ː

* Significant interaction of Pen and Original label in CISP 1a-c
* Signiticant interaction of Vowel (lip rounding) and Original label in Kang et al. (2016).

From learningː

* One-sided blocking in LJ18 consistent with pen as plausible alternative explanation for shifted /ʃ/’s but not shifted /s/’s
* Chart

  Description automatically generatedPotentially in line with our results, but need to run CISP-1 vs CISP-2 model

WHYː

* Is lip-rounding only a salient cue for /ʃ/?
* Is it about precision?
* Do we even need to determine this?

2.4 Discussion

Experiment 1 investigated whether presence of a pen in a talker’s mouth affects listeners’ perception of the same acoustic stimuli along an asi-“ashi” continuum. This is indeed the case. Specifically, the presence of a pen was S-biasing: the same acoustic input paired with video of the pen in the hand resulted in more “ashi” responses, whereas the pen in the mouth resulted in more asi responses.

This is unexpected if visually presented causes were ignored by speech perception. Rather, our result suggest that listeners might expect the pen to cause pronunciations to sound more “ashi”-like. The decreased proportion of “ashi” responses would then reflect compensation for this expectation. (a number of alternative explanations are ruled out by the fact that the effect of the pen is independent of whether the video recording was of an asi or “ashi”). Our findings thus present initial evidence that listeners take into account how incidental causes might affect pronunciation. Finally, the robust effect of the acoustic continuum suggests that our audiovisual stimuli successfully replicate previous findings from audio-only tests. This is critical for Experiment 2, which employs the exact same test stimuli and procedure as Experiment 1b while manipulating prior exposure to the test talker between subjects.

Our results are compatible with a number of different explanations, all of which point either to the importance of visual cues to articulation in the perception of speech or to causal inference. First, it is possible that the pen visually occludes critical visual information about /s/ vs. /ʃ/. Specifically, /ʃ/ involves lip-rounding whereas /s/ does not. It is possible that the pen’s location in the mouth occludes the lip-rounding. If this is the case, then how do listeners react to the absence of this information? One possibility is that listeners expect a pen in the mouth to prevent lip rounding. In that case, this absence is no longer a cue to /s/, and only videos that were originally taken from /s/ words are expected to be affected by the pen. Alternatively, if listeners do not expect the pen to affect lip rounding, they might simply not perceive the feature that would usually mark an /ʃ/, and would now categorize those sounds as /s/. In this case, we would expect only videos that were originally taken from an /ʃ/ word to be affected by the pen. Finally, listeners might expect the pen to change the acoustic realization of the token towards /ʃ/ and compensate for it. This would explain why we see an effect of the pen independent of the visual labels. Finally, it is worth noting that if the pen were to erroneously create the impression of lip rounding, we would expect a different result altogether.

Experiments 1a-c also serve as a basic validation of the test phase used in the perceptual recalibration experiments presented below. Those perceptual recalibration experiments consist of an exposure and test phase, both of which employ audiovisual stimuli. While some previous experiments have employed audiovisual stimuli during exposure, most previous perceptual recalibration experiments have exclusively use audio-only stimuli during the test phase (but see Lüttke et al., 2018).

As predicted by compensation for co-articulation, visual presence of a pen in the mouth (ostensibly shifting acoustics towards the “sh” category) caused listeners’ perceptions of identical acoustics to be biased towards the “s” category. Equipped with knowledge of this effect, a review and re-evaluation of work in adaptation including a pen in the mouth is warranted.

# Experiments 2 & 3: Adaptation

This literature has employed the paradigm of lexically guided perceptual learning, a common methodology for measuring how exposure to atypicality in production from a novel talker results in altered categorization of subsequent speech from that talker. The paradigm commonly employs two phases, being exposure and test. The test phase is traditionally an audio-only instantiation of the methods reported above as Experiment 1.

ADD: general overview of Exposure paradigm in LGPL

The perceptual recalibration effect is generally robust and does not seem to require much or any conscious attention; subjects adapt even given an environment full of distractions, other cognitive load-bearers, or explicit instruction that the speaker or tokens are unreliable or altered (Drouin & Theodore, 2018; Liu & Jaeger, 2019). As few as 4 shifted tokens has been found to elicit significant learning (Cummings & Theodore, 2023).

However, despite the effect’s rapid and robust nature, it does not appear indiscriminate or unable to be blocked. Kraljic, Samuel, & Brennan (2008, Experiment 1b) exposed listeners to 10 shifted tokens which co-occurred with video including a pen in the mouth of the talker. These were followed by 10 unshifted tokens co-occurring with video of the talker without the pen in her mouth. This second section of exposure served to allow listeners to infer that the shift in production was in fact resultant of the pen; productions became normal as soon as the pen was removed. In subsequent audio-only test, participant groups between biasing conditions did not show a shifted /s/-/ʃ/ categorizations from one another. This is taken as support that the presence of a plausible alternative explanation for an altered pronunciation blocks perceptual recalibration. Specifically, the authors posit that “The system integrates available cues about whether a variation is characteristic of the speaker who is producing it or an incidental consequence of some other factor. If the variation seems characteristic, the appropriate phonemic representation is restructured to accommodate it; if the variation seems incidental, no such restructuring occurs.” (Kraljic, Samuel, & Brennan, 2008).

The results and interpretation of Kraljic, Samuel, & Brennan (2008) leave open at least two alternatives. The most succinct is perhaps a framework wherein all audio input that co-occurs with atypical visual input (such as a pen in the mouth or any other incidental cause) is discarded entirely by the listener. This postulates that perceptual adaptation is driven only by prior perception in typical contexts and should only be seen experimentally when exposure includes these contexts. This hypothesis parsimoniously explains the lack of learning but is unlikely given the finding of the current Experiment 1 that visual evidence of incidental causes is accounted for in perception. An alternative is that phonetic restructuring may be specific to the environment in which it occurs, a hypothesis furthered below.

In a follow-up study, Kraljic & Samuel (2011, Experiment 2) found that learning is preserved when shifted pronunciations continue in the second section of exposure (wherein the pen is no longer in the talker’s mouth). This remained the case when visual evidence was entirely absent from the second section of exposure, which importantly suggests an assumption by listeners that speech evidence in absence of video is produced without an incidental cause. Learning was also found when a talker’s speech was atypical only during the second section of exposure, but normal when the pen was in the talker’s mouth (Kraljic & Samuel, 2011, Experiment 3). These findings are taken as evidence that the effect of the pen on learning (or lack thereof) follows from the way humans store and retrieve previously experienced speech input. Specifically, it’s well documented that speech input is stored along with the context in which it occurs, and that memory traces of speech ‘exemplars’ or ‘episodes’ are used to categorize subsequent input (Johnson, 1997; Pierrehumbert, 2002; Kleinschmidt & Jaeger, 2015). Under this account, perceptual adaptation is driven by the linked storage of auditory and visual (typical or otherwise) input. Retrieval of these memory traces should result in perception of subsequent auditory input being colored by previous experience within the same visual context.

Liu & Jaeger (2018, Experiment 1b) provides the first study wherein causal ambiguity between the talker and the pen was not disambiguated during exposure. Shifted sounds co-occurred with video of the talker with the pen in her mouth for the entirety of exposure. In subsequent audio-only test, learning was found to be blocked only for the “sh”-labelled group (those hearing words ambiguous between, e.g. publisher and publiser). This provides the first evidence for an effect of compensation in exposure towards learning exhibited in test. Liu & Jaeger (2018) identify that this asymmetry “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”.

These studies collectively, while providing a wealth of intriguing findings, are all crucially limited by their audio-only testing phases. If learning is indeed linked to retrieval of previously stored audio-visual exemplars, then the modality shift from visually informed exposure to non-visually informed test is expected to require generalization and constrain learning. Robust support of context-specific learning cannot be found when context changes between the phase of the experiment designed to elicit learning and the phase which measures it. Finally, a crucial test of context-specific learning is to determine whether shifts encountered in context of the pen in the mouth, which have been found not to elicit adaptation in audio-only context, may still elicit adaptation in subsequent matching context (being, test to percepts including the pen in the mouth).

Experiment 2 of the present study therefore seeks to replicate Experiment 1b of Liu & Jaeger (2018), with the addition of visual information in test.

Experiment 2

Having established that the visual presence of an incidental cause affects listeners’ perception of an unfamiliar talker’s speech, prior to any informative exposure to the talker, our second goal was to examine how this alteration in perception manifests itself in adaptation. To this end, we ran three exposure-test Experiments (2a-c). These experiments employ the exact same design and procedure but differ in the specific visual and acoustic stimuli they employ, as well as minor details of the post-experiment survey, and one significant shift in the pre-experiment practice phase. Specific differences are described in the methods.

Participants first underwent an exposure phase, wherein they were presented audiovisual speech stimuli of both English words and nonwords. A lexical decision task was used to ensure attention. Crucially, words typically containing either /s/ or /ʃ/ (split between-participant groups) were shifted in this phase to contain instead an ambiguous sound between /s/ and /ʃ/. During production of the critical /s/-/ʃ/ fricative, the talker either had the pen in the mouth (Figure 1, Panel A) or in the hand (Panel B), again split between-participant groups. This exposure phase was paradigmatically identical to that of Liu & Jaeger (2018; Experiment 1).

All subjects then underwent a test phase identical to that of Experiment 1c.

## Methods

Except for the use of audiovisual rather than audio-only test stimuli and minor procedural changes reported below, Experiments 2a-c closely followed Experiment 1a in Liu & Jaeger (2018).

***Participants.*** Prolific things, to be added

Participants were excluded from the analysis if XYZ

***Materials.*** The exposure phase used identical stimuli to that of Liu & Jaeger (2018, Exp. 1a). The test phase used identical stimuli to those of Experiment 1c.

***Procedure.*** Our web-based procedure closely followed Liu and Jaeger (2018). We used the same HTML, CSS, and Javascript code (with the modifications necessary to accommodate our design) and same recruitment language for the HIT. The experiment thus largely looked identical to Liu and Jaeger (2018). The recruitment screen and instructions were identical to Liu and Jaeger (2018). The experiment consisted of (1) instructions, followed by (2) an exposure phase, (3) a test phase, and (4) a post-experiment survey. An exact replica of Experiment 2 for demonstration purposes can be found at XYZ. The same page contains screen recordings of a participant taking the experiment.

3.2 Procedure

*Instructions.* ***.*** Prolific things, to be added

*Exposure*

Although not reported in Liu and Jaeger (2018), the experiment employed catch trials during exposure. On those catch trials, a white dot appeared in the video. The dot appeared roughly in the center of the video, close to the talker's mouth---the region of the video that indicates the presence of an alternative cause for unexpected pronunciations (the pen in the mouth). Participants were instructed to press "B" when they noticed the white dot, and that this task served to distinguish human participants from robots (catch trials with similar language are not uncommon on Mechanical Turk). These instructions were first provided for the practice block, and remained in place during both the exposure and test block. One of the two videos during the practice phase contained a white dot, as did 12 videos during the exposure phase: six catch trials occurred with filler word videos and six occurred on filler non-word videos. Because of changes made to the list generation (described below), we use 10 of these catch trials: half on filler word trials and half on filler non-word trials. On half of these catch trials the speaker had the pen in the mouth, on the other half the pen was in the hand.

We also realized that having the participants press X or M after pressing B created some confusion: they had to wait before clicking X or M, or else sometimes the exposure phase would not proceed. Thus, unlike Liu and Jaeger (2018), the catch trials in the present experiment end when "B" is pressed whereas participants in Liu and Jaeger (2018) had to \*additionally\* answer the word vs. non-word question. We removed the latter task from the catch trials since our piloting showed that this trial structure---the need to press two buttons on some trials---was confusing. As a consequence, the lexical decision accuracies and reaction times we report below for the exposure phase are based on 88, rather than all 100, trials per participant. This change in method is not predicted to change the results. It does, however, provide us with an additional measure of the extent to which participants pay attention to the visual stimulus, on which causal inferences are hypothesized to be contingent.

Following LJ18, we use 100 exposure trials, consisting of 10 words with typical fricatives, 10 words with shifted fricatives, 30 filler words, and 50 filler non-words. This is the same proportion of typical, shifted, and filler trials used in other previous studies (Kraljic et al., 2008; Kraljic & Samuel, 2011). LJ18 generated two pseudo-randomized stimulus orders for exposure. These lists and their respective reverse orders formed four different exposure orders. Key binding---whether "X" and "M" corresponds to a word response or a non-word response, respectively, or vice versa---were counterbalanced within each of these four orders, resulting in a total of 8 different exposure lists. The two pseudorandom orders in LJ18 were generated by repeatedly randomizing the order of stimuli until critical items, fillers, and catch trials were somewhat evenly distributed across the list. We take a more systematic approach with the goal to create many randomized stimulus orders across participants, thereby creating cross-participant variability in nuisance factors not expected to affect test performance. Stimulus order was generated online, as part of the experiment. Invisible to participants, we divided exposure into ten blocks, each consisting of one word with a typical fricative (/s/ or /sh/ depending on the exposure condition), one word with shifted fricative (/sh/ or /s/ depending on the exposure condition), three filler words, and five filler non-words. The assignment of the specific videos to each block---for example, which of the 10 words with a typical fricative was assigned to block 1---was randomized. The order of the 10 stimuli in each block was also randomized under the constraint that two typical or two shifted words could never occur in adjacent list positions. The exposure list was obtained by concatenating the ten blocks, resulting in the 100 exposure trials.

Catch trials were evenly distributed across the ten blocks, with one catch trial per block. For each block, a random coin flip determined whether the catch trial was on a filler word or non-word. Across the 10 blocks, catch trials occurred on exactly 5 filler words and 5 filler non-words. Half of the catch trials occurred with the pen in the mouth and half with the pen in the hand.

*Test*

As in Liu and Jaeger (2018), participants listened to nonce-word stimuli drawn from the asi-ashi continuum, and were instructed to respond "whether the speaker is saying “asi” or “ashi" (and to press "B" when they saw a white dot, which never happened during test). These categorization responses form the primary data to assess perceptual recalibration and test our hypothesis.

The 72 audio-visual stimuli were grouped into six blocks of 12 stimuli, so that each block fully crossed the two video conditions (pen in hand vs. mouth) with the six audio conditions (step along the asi-ashi continuum). We further made sure that the same video was never repeated within the same block, and that each block contained exactly six videos that had been extracted from a video with a word containing /s/ (3 with the pen in the hand, 3 with the pen in the mouth) and six videos that had been extracted from a video with a word containing /sh/ (3 with the pen in the hand, 3 with the pen in the mouth). Following Liu and Jaeger (2018), stimuli were randomized for each participant while obeying the constraints described here.

*Post-experiment survey*

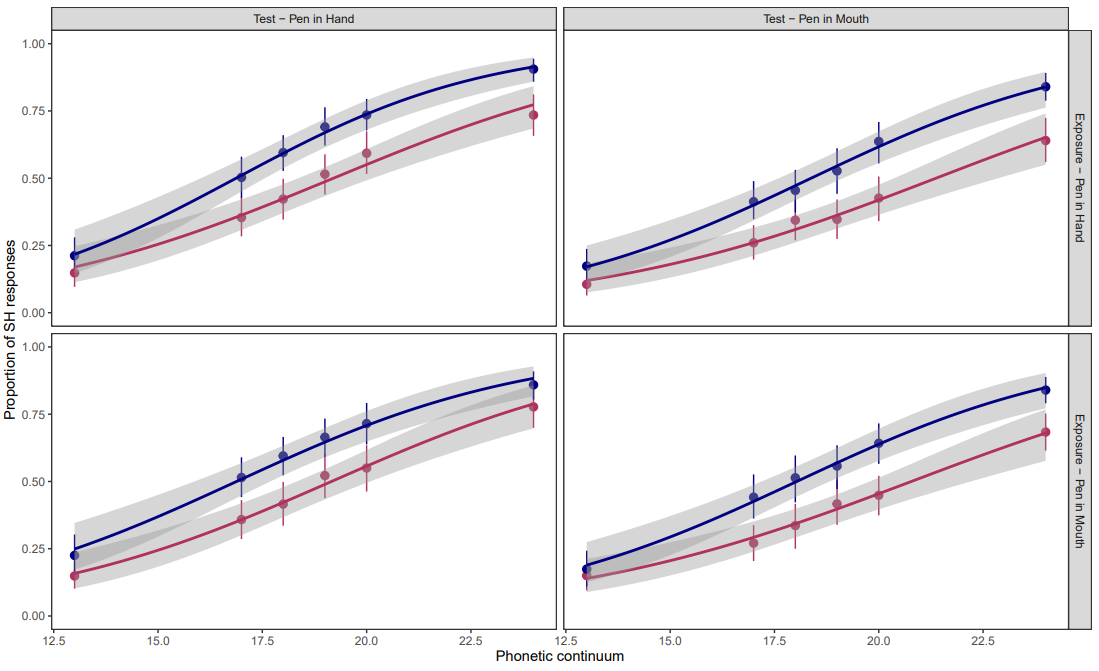
Following Experiment 1b, we made minor changes to the survey, with the goal to increase the clarity of our questions about internet connectivity. Specifically, we changed the two questions that asked about video or sound issues during the experiment. We clarified that these questions were asking solely about technical issues, rather than any oddities between the alignment of the video and audio (as separate questions already assessed that aspect). The revised questions were introduced by the following statement:

"The videos and sounds in this experiment were manipulated by aligning the same video with different sound sources. As a consequence, you might have noticed some 'jumps' or slightly odd looking moments in the video. \*\*Here we are interested in \*potential technical issues with the internet connection\* beyond any oddities you might have noticed about how the video and sound aligned.\*\*" (emphasis in original)

Then two questions assessed whether the videos or sounds failed to load or stopped and restarted playing during exposure. As these revised questions explicitly mentioned that the videos and sounds had been manipulated, they were moved towards the end of the survey, so that they followed any questions that inquired about the pronunciation of the talker in the video (recall that participants could not go back to previously answered questions).

3.2.4 Exclusion criteria

We followed the same exclusion criteria as for Experiments 1a and 1b, including excluding participants who did not make the acceptance criteria but who took the experiment anyway. Additionally, we excluded participants with a less than 0.85 lexical decision accuracy in exposure. Following Liu and Jaeger (2018), we OTHER EXCLUSIONS

3.3 Results

* (equally) robust learning in all conditions

Discussion:

At first blush, robust learning in all conditions seems counter to theories of context-specific learning. Important to consider here, however, is the overall contextual similarity between exposure and test for the current study. Prior work has established that evidence of shift provided in visual pen-in-mouth episodes does not seem to generalize to audio only test. Two crucial shifts occur herein: causality and modality. Recall the 2nd experiment of Kraljic & Samuel, 2011, establishing that audio-only tokens are assumed to be characteristic of the talker, given no evidence to the contrary. Visual evidence without incidental cause, in contrast, is able to generalize to audio only test, as causality is held constant through the change in modality. The conditions of the current Experiment 2 always maintain modality across exposure and test, and therefore require generalization across at most the single dimension of causality. Recall additionally, via the results of Liu & Jaeger (2018, 1b), that unresolved causal ambiguity only blocked one of two learning conditions. This provides plausibility to at least some generalization across cases where causality is ambiguous. Experiment 2, wherein modality never changes and causality is never strongly disambiguated in either direction, appears to meet criterion for generalization of learning in all four conditions.

This explanation, reliant on at least partial generalization across ambiguous contexts, makes a strong prediction for a case wherein modality is held constant and causality is disambiguated in exposure. Specifically, while some amount of learning may occur across conditions, more learning is expected in cases where the disambiguated causality of exposure matches the presented context of test. Experiment 3 examines this hypothesis via an audiovisual replication of Kralic, Samuel, & Brennan (2008, Experiment 1b). Subjects first experience shifted tokens in context of a pen in the mouth, then have causality disambiguated through a second section of exposure with unshifted tokens and the pen no longer in the mouth. Test is then provided to both pen in mouth and pen in hand percepts, with context-specific learning predicting more evidence of adaptation present for pen in mouth test.

# Experiment 3

methods:

results:

Discussion:

HYPOTHESIS TREE:

1. PIM percepts ignored:

* Could explain E2, but really unlikely based on LJ18 and our E1 & E3

1. Pre-categorization normalization (Causal CfC) drives the effects; PIM has no effect on adaptation but just on perception:

* Works for E1
* E2:
  + Exposure Predictionsː Causal compensation predicts all PIM tokens to behave as if they are higher spectral CoG than they really are. In s-bias (red) conditions, this could inhibit adaptation; in sh-bias (blue) conditions, it is expected either to have no effect or to increase adaptation. The ‘double’ shift of the acoustic shift and compensation being directionally the same in sh-bias conditions could potentially move those tokens out of the ‘goldilocks’ zone (Babel et al. 2019), but this is a priori unlikely and doesn’t pattern out in the data. Since the effect of the pen is handled through in the moment compensation, and no evidence is provided that e.g. the talker *only* exhibits a shift while PIM, causal compensation does not expect to block learning (as a difference between groups).
  + NOTEː Causal compensation predicts that s-bias conditions should not induce adaptation (as there isn’t a difference between expectations and results, while ʃ-bias conditions should adapt. This is the exact opposite of what Liu & Jaeger (2018; Exp 1) found...
  + Test Predictionsː As in Experiment 1, PIM tokens should be lower in /ʃ/ responses than their PIH counterparts. Learning (as a difference from baseline/exp1) should be exhibited by the ʃ-groups moreso than the s-groups.
  + Resultsː Overall patterning is that all conditions learn about the same amount. For PIH test, s- vs ʃ-bias conditions seem to move about the same distance from baseline, rather than the s-bias staying put (as most predicted by the Causal Compensation hypothesis), but neither condition significantly differs from baseline. For PIM test, the s-bias condition does seem visually to align more with the PIM test in Experiment 1, but again no differences from baseline are significant. Overall, no glaring issues with a causal compensation hypothesisǃ
* E3:
  + First Exposure predictionsː Same as Experiment 2; learning in sh-bias but not s-bias groups.
  + Second Exposure predictionsː Causality is disambiguatedǃ For sh-bias group, perhaps ‘unlearning’ is predicted, but perhaps not (see Cummings & Theodore, 2022; consistency of input doesn’t seem to matter as much as quantity). For s-bias group, causal compensation gains support (as it predicts the pattern of Amb-PIM to clear-PIH). IF uncertainty maintenance were at play to begin with (and perhaps listeners were only compensating for a portion of lip rounding’s expected effect), then perhaps this support will strengthen the effect.
  + Test predicitonsː If causal compensation becomes more robust (resultant of its confirmation via exposure 2), then the s-bias condition in PIM test is expected to have even fewer /ʃ/ responses than experiment 2.
  + Resultsː Both conditions learn, but the PIM condition exhibits significantly *more* learning (driven by the s-bias condition). While the paragraphs above explain how Causal Compensation *could* predict this pattern, it’s admittedly tenous and HARK-ey...

1. Pre-categorization happens during perception, but percepts are ADDITIONALLY stored along with their context:
   1. Passive context-based storage/retrieval
   2. More sophisticated causal reasoning

# Acknowledgments

We are deeply grateful to Linda Liu for answering our questions about the stimuli and procedure employed in Liu and Jaeger (2018), and to Zach Burchill for invaluable technical support with several Javascript issues. We also thank Zach Burchill, Wednesday Bushong, Linda Liu, and Xin Xie for sharing materials and feedback for a tutorial on conducting crowdsourcing experiments via Mechanical Turk that was developed as part of this project (available via [https://github.com/hlplab/Tutorial-MTurk-experiments-via-mturkutils]). We thank Dr. Arty Samuel for generously providing the audio stimuli, and Dr. Molly Babel and Jamie for generously providing the video stimuli used in this experiment. Finally, we thank Dr. Chigusa Kurumada, Dr. Ralf Haefner, and the BCSC 206/207 class of 2020-2021 for helpful discussion and feedback on this project.

# Appendices

*A.* All HTML, Javascript, style sheets, and materials of the experiments are shared as part of the OSF repository (see experiments/). 5\*This includes fully executable versions of the experiment that can also be viewed at (https://www.hlp.rochester.edu/mturk/CISP/experiment-A/experiment-A.html?URL\_PARAMS), where the necessary URL parameters are described as part of the YAML files on OSF (see experiments/YAML files for MTurk/). For example, an instance of Experiment 1a can be called by (https://www.hlp.rochester.edu/mturk/CISP/experiment-A/experiment-A.html?label=no\_exposure&respKeyTest=0&testSet=A). URL parameters were not visible to participants. The post-experiment survey for each experiment is available as part of the OSF repository (see experiments/experiment-NAME/surveys/)

*B.* Participants were recruited by advertising a "Speech Perception Experiment. MUST BE NATIVE ENGLISH SPEAKER AND WEAR HEADPHONES. MUST USE CHROME." along with the brief description "Listen to sentences. Takes approximately 5-10 minutes." The only additional information available to participants in deciding whether to accept this HIT (human intelligence task) was the general description provided on the first page of the instructions.

*C.* An instance of Experiment 2 can be called by (https://www.hlp.rochester.edu/mturk/CISP/experiment-A/experiment-A.html?label=S&condition=M&respKeyExp=0&respKeyTest=0&testSet=B).

# References

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

Bejjanki, V. R., Beck, J. M., Lu, Z. L., & Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience*, *14*(5), 642–650. https://doi.org/10.1038/nn.2796

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

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

Drouin, J. R., & Theodore, R. M. (2018). Lexically guided perceptual learning is robust to task-based changes in listening strategy. *The Journal of the Acoustical Society of America*, *144*(2), 1089–1099.

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

Franken, M. K., Eisner, F., Schoffelen, J., Acheson, D. J., Hagoort, P., & Mcqueen, J. M. (2017). Audiovisual recalibration of vowel categories. *Interspeech*, 655–658.

Klatt, D. H. (1986). The problem of variability in speech recognition and in models of speech perception. Invariance and Variability in Speech Processes, 300–319.

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

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*(3), 459–465.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability: Research article. *Psychological Science*, *19*(4), 332–338.

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

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174* (June 2017), 55– 70.

Liu, L., & Jaeger, T. F. (2020). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *45*(12), 1562–1588.

Lüttke, C. S., Pérez-Bellido, A., & de Lange, F. P. (2018). Rapid recalibration of speech perception after experiencing the McGurk illusion. *Royal Society Open Science*, *5*(3). https://doi.org/10.1098/rsos.170909

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices (McGurk Effect). *Nature*, *264*(5588), 746–748. https://www.naturecom.libproxy1.usc.edu/articles/264746a0.pdf%0Ahttps://www.nature.com/articles/264746a0.pdf

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

Proctor, M., Shadle, C., & Iskarous, K. (2006, December). An MRI study of vocalic context effects and lip rounding in the production of English sibilants. In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 307-312).

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

Xie, X., Liu, L., & Jaeger, T. F. (2021, January 11). Xie, Liu, & Jaeger (2020). Cross-talker generalization during foreign-accented speech perception. https://doi.org/10.1037/xge0001039

1. This differs somewhat from the approach taken in Liu and Jaeger (2018), which did not require *significant* reversal of the slope for participant exclusion. We consider the approach taken here more conservative, and decided on this change prior to analysis. [↑](#footnote-ref-2)