Compensation in audiovisual speech perception: discounting the pen in the mouth

Shawn N. Cummings[1,3,4,5], Gevher E. Karboga[1], Menghan Yang[1,4], and T. Florian Jaeger[1,2]

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

[2]Goergen Institute for Data Science and AI, University of Rochester, USA

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

[4]Department of Psychological Sciences, University of Connecticut, USA

[5]Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, USA

Word count (main text): ~9200

Contact author:

Shawn N. Cummings
shawn.cummings@uconn.edu
David C. Phillips Communication Sciences Building
Storrs, CT 06269
USA

Abstract

The articulation of speech segments is influenced by their phonetic context. Human speech perception seems to compensate for such coarticulatory effects, interpreting acoustic cues relative to their values expected in the current context. We test whether similar compensation is observed for visually presented, non-phonetic contexts (a pen in the mouth of a talker), as predicted by some accounts of perceptual compensation. In a series of perception experiments, we find that listeners compensate for the presence of a pen in the mouth of the talker, as long as the effects of the pen on the articulators (e.g., lip shape) are visually evident. Beyond demonstrating perceptual compensation for non-phonetic contexts, these findings also inform ongoing theoretical debates in the literature on perceptual recalibration, where similar manipulations have been found to block or reduce perceptual learning.

Keywords: speech perception; audiovisual; compensation; coarticulation

Introduction

The phonetic realization of sound categories is affected by their phonetic context, a process known as coarticulation. For example, English fricatives have a lower spectral center of gravity directly following the vowel /u/ (as in *moon*) compared to front vowels (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. Speech perception is known to *compensate* for such coarticulatory effects on production: 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 the lowered spectral center of gravity at least in part to the coarticulatory effect of the preceding /u/, rather than an intention to produce a /ʃ/ (Fowler, 2006). Similar compensation has been documented for a wide range of acoustic, or phonetic contexts, sometimes under the alternative term *normalization* (e.g., Cole et al., 2010; Francis et al., 2006; Holt Huang & Holt, 2009; McMurray & Jongman, 2011, 2016; Syrdal & Gopal, 1986; for review, see Weatherholtz & Jaeger, 2016).

There is some evidence that compensation is not limited to acoustically conveyed contexts. For example, in an effect resembling that of preceding /u/, visually presented lip-rounding—which tends to be correlated with lowering of the third formant (F3)—immediately preceding audio of an ambiguous /d-g/ blend biases listeners towards perceiving /g/ (Fowler et al., 2000; Kang et al., 2016; Mitterer, 2006). 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 preceding visual context of lip-rounding. Results like these led to the hypothesis that compensation can occur regardless of the type and modality of contextual cues. As Fowler (2006, p. 166) put it: 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 [for reasons other than the production of the /d/-/g/ sound] and, therefore, would lower the F3 of the syllable that the gesture overlapped with temporally" (Fowler, 2006).

This conjecture has not stood unchallenged (for a concise and balanced review, see Viswanathan & Stephens, 2016). For example, some studies failed to replicate the effect of visually presented context (Vroomen & de Gelder, 2010), or only replicated it under certain conditions (Holt, Stephens, & Lotto, 2005).[1] Most strikingly though, previous studies have—to the best of our knowledge—exclusively focused on visually presented *phonetic context*. In these studies, participants watch a video of a talker's face articulating speech (e.g. for the word *alda* or *arga*) while hearing audio that may or may not be the result of those articulatory movements (e.g., some acoustic blend of *alda* and *arga* might be dubbed onto the video for *alda*). These studies investigate whether visual evidence of the phonetic context (e.g., the /l/ or /ɹ/ in *alda* or *arga*) affects the perception of the target sound (e.g., the /d/ or /g/). However, if it indeed does not matter whether the talker is "about to whistle a merry tune or about to kiss a loved one", then even *non*-linguistic visual context should elicit compensation similar to that observed in previous studies on linguistic context (for related discussion, see Fowler, 2004; Rosenblum et al., 2016). This is the prediction we test here.

Specifically, we simulate a talker whose articulation is organically affected, but crucially by *non*-phonetic context. Consider a talker with a pen in the mouth (as in Figure 1 below), producing an /s/ or /ʃ/. A pen in the mouth has two visually evident effects on articulation. The first 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

[1] Effects of visually presented context seem to be strongest when the relevant visual evidence is particularly clear and still present during the articulation of the target sound (see discussion in Fowler, 2006; Lotto & Holt, 2006). We return to this in Experiment 2.

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 (McFarland & Baum, 1995). Lip rounding is similarly expected to lower the spectral center of gravity for surrounding fricatives by effectively temporarily increasing the length of the vocal tract (Lindblom & Sundberg, 1971). As lower spectral center of gravity is one of the primary cues distinguishing /ʃ/ from /s/ in English (Jongman, Wayland, & Wong, 2000), 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, this compensation should bias their perception towards /s/ (against /ʃ/), relative to an identical acoustic input that is observed in the absence of a pen in the mouth. We test this hypothesis in a series of web-based experiments on audiovisual speech perception. Experiments 1a-c demonstrate the basic effect of interest. Experiment 2 begins to elucidate the necessary conditions for the effect.

As we describe in more detail after presenting our results, the experiments we present here also speak to an ongoing debate in a separate line of research on perceptual recalibration. In perceptual recalibration experiments, listeners are exposed to speech from an unfamiliar talker who pronounces a particular sound category in an unexpected way (e.g., Kraljic & Samuel, 2006; Norris et al., 2003). For a particular group of participants, the talker might, for example, pronounce all /s/ sounds in a way that make them sound rather /ʃ/-like (e.g., using pronunciations like *dinoshaur, medishine, etc.*, in Kraljic and Samuel, 2006). Following such exposure, listeners tend to hear more tokens along an audio-only /s/-to-/ʃ/ test continuum as "s", suggesting that exposure shifted or expanded listeners' "s" category (Kleinschmidt & Jaeger, 2015; Cummings & Theodore, 2023).

One influential finding in this line of work is of direct relevance to the present study: perceptual recalibration to /s/ or /ʃ/ has been found to be partially or completely blocked if the talker has a pen in the mouth during the pronunciation of the shifted words (Kraljic et al., 2008; Kraljic & Samuel, 2011; Liu & Jaeger, 2018). Why this happens has remained a matter of theoretical debate,

and competing proposals appeal to different learning and memory mechanisms. For instance, one hypothesis holds that listeners store audio-visual exemplars of the talker with the pen in the mouth separately from audio-visual exemplars with the pen in the hand, and that only the latter type of exemplars affects listeners' responses during the audio-only test phase (Kraljic & Samuel, 2011). This is where the present study potentially becomes relevant: if the pen in the mouth affects *perception* (through compensation)—as we test here—this would potentially preempt the need to appeal to *learning* or *memory* to explain the blocking of recalibration. We return to this point in the general discussion, where we also clarify why the design of perceptual recalibration experiments is not suited (and, of course, was never meant) to address questions about compensation.

Open science statement

All experimental materials—including the original video and audio recordings as well as all audiovisual test stimuli for all experiments along with their phonetic annotations—lists, and trial-level data are available as part of the OSF repository at https://osf.io/2asgw/. The same holds for the JavaScript code for the experiments, and the R code for analyses and visualizations. The latter is made available in the form of a "knittable" R Markdown document that generates the supplementary information for this article through a single click in a freely available software (*R*, R Core team, 2023; *RStudio*, Posit team, 2024). Exact replica of all experiments for demonstration purposes are available at https://sites.google.com/view/causal-inference-in-speech/urls-to-our-experiment.

The experiments we present here were *not* pre-registered via OSF. They were, however, pre-registered—and conducted—in the context of an undergraduate research class in the Brain and Cognitive Sciences at the University of Rochester. The five experiments presented here were conducted as part of a larger project to distinguish between competing explanations for the blocking of perceptual recalibration when the talker has a pen in the mouth during the shifted pronunciations

(e.g., Kraljic et al., 2008), as described above. This larger project seeks to understand how such incidental causes affect listeners' interpretation of the acoustic input (1) in the moment ('processing' / 'perception')—the question addressed here—and (2) beyond the moment during processing of subsequent input from the same talker ('adaptation' / 'perceptual learning'; building on the seminal work of Kraljic et al., 2008). Where our design decisions for the present work were motivated by the goal to ultimately also address question (2), we mention so below.

Aggregate Demographic Information About Participants

Because the demographic composition of our participants did not vary significantly across experiments, we report aggregate information here. All demographic categories were based verbatim on National Institutes of Health (NIH) reporting requirements. Across all five experiments, 47.0% of our participants reported as female, 52.0% as male, and 0.9% declined to report gender. The mean age of our participants was 36.9 years, with an interquartile range of 28–44 years (SD = 12.1; 1.6% declined to report). All participants reported to be at least 18 years of age. With regard to ethnicity, 8.5% of the participants reported as Hispanic, 89.7% as Non-Hispanic, and 1.9% declined to report. With regard to race, 72.7% reported as White, 12.5% as Black or African American, 7.8% as Asian, 3.1% as More than one race, 0% as American Indian/Alaska Native or Native Hawaiian or other Pacific Islander, 0.6% as other, and 3.1% declined to report. As we have no theoretical reasons to investigate demographic effects on the outcomes reported in the present study, we refrained from doing so.

Experiments 1a-c

Experiments 1a-c test how the presence of a pen in a talker's mouth affects listeners' interpretation of that talker's speech. 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 (see Methods). Participants were presented with 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. During the production of the critical /s/-/ʃ/ fricative, the talker held a pen either 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 evident effects on the articulation of /s/ and /ʃ/—affects the interpretation of acoustic cues to the /s/-/ʃ/ contrast.[2] Participants performed a two-alternative forced choice identification (categorization) task, answering whether they thought the talker in the video said *ashi* or *asi*.

Figure 1 Illustrating the critical manipulation in Experiments 1a-c. Participants saw and heard audiovisually 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). ***Note that actual stimuli in Experiment 1a-c (available at osf.io/2asgw) used a female talker.*** *As we do not have permission to publish images of the talker (only to use them in the experiment), we present here a mock-up of the manipulation (consent for the use of the stimuli outside of publication was confirmed).*

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

[2] Note that this is a different manipulation, than the one in previous research on perceptual recalibration described in the introduction. Those studies assessed how listeners' categorization responses during an *audio-only* test phase were affected by a preceding audiovisual *exposure* phase that manipulated pen placement (no categorization responses were elicited during exposure). Such designs leave open how, and why, perception is affected *in the moment* by the placement of the pen, which is the question we address here.

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/ affects 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.[3] Instead, 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

[3] Only a few previous studies have gradiently manipulated visual cues to articulation (e.g., Bejjanki et al., 2011; Kang, Johnson, & Finley, 2016). The studies have employed either obvious animation, or a single 'ambiguous' real-life video. None of these studies modeled the visual consequences of an articulatory obstruction (like a pen in the mouth).

changes reported below, Experiments 1a-c closely followed the norming experiments in Liu & Jaeger (2018). All participants were recruited under Protocol 00045955 approved by the Research Subjects Review Board at the University of Rochester.

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 counter-balanced nuisance variables described below. Participants took an average of 22.6 minutes to complete the experiment (SD = 18.5 minutes) and were remunerated $6.00/hour. Participant exclusions never exceeded 10% and are reported in Table 1, discussed below.

Participants 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, (iii) met the software requirements (a recent version of the Chrome browser engine), and (iv) had not previously participated in any similar experiments from our lab. Before the experiment could be accepted, participants had to confirm that they were (v) native speakers of US English, (vi) in a quiet room without distractions, (vii) wearing over-the-ear headphones.

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 in Kraljic et al. (2008) and employed in many subsequent studies since. Following previous work, we selected six steps along the 31-step continuum. 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 statistical power to detect other effects—including the hypothesized effect of pen location—is highest at test steps that elicit close to 50% *ashi* and 50% *asi* responses.

Following experiments on perceptual recalibration, we thus aimed to select one continuum step that, across all other manipulations, yields approximately 25% *ashi*-responses, four steps that yield close to 50% *ashi*-responses, and one step that yields 75% *ashi*-responses (e.g., Kraljic et al., 2008 and later work).

This goal resulted in three very similar experiments (Experiments 1a-c), which differ only in the six selected acoustic continuum steps, as well as some aspects of the exit survey described later (for details, see SI). When we present the results from each experiment below, we include visualizations of the continuum steps that were included. This demonstrates that Experiment 1c achieved the intended placement of continuum steps, while also allowing us to test whether the effects of the pen are robust to the specific continuum steps chosen (i.e., whether the effects held across 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 by Babel (2016) because the original video stimuli from Kraljic et al. (2008) 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), providing a highly plausible match for the voice of the talker in audio recordings (as confirmed in Babel et al., 2016; 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. on July 17, 2020). For Experiment 1a, we thus identified exposure videos with the required sound sequence similar to *ashi* (e.g., *m[achi]nery*) or *asi* (e.g., *democr[acy]*). Only the twelve videos in which this sequence was of very similar duration as the *ashi-asi* nonce-word recordings were used (see SI, for full list). We used the open-source video editing software Shotcut (shotcut.org) to extract the relevant video frames from the original recordings. Following the procedure used by Babel to create the exposure videos, 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, so that the presence of a visual bias towards /s/ or /ʃ/ and the location of the pen were 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 the results of Experiment 1a indicated a particularly strong visual bias for that video.

Audiovisual 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.

Procedure. The experiment consisted of (1) instructions, followed by (2) a test phase and (3) an exit survey.

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 experiment, but in order to start the experiment, participants had to

accept the experiment.

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 planned subsequent experiments on question (2) mentioned in the Open Science Statement. None of the trials during the test phase actually contained a white dot. Participants then completed 72 trials of a 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. 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.

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 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.

Exit survey. The survey for Experiment 1a was identical to that of Liu and Jaeger (2018). All questions are listed in the SI. Questions assessed the quality of the audio equipment and whether participants experienced stalling of audio or video (to help us catch code problems). The survey also contained a catch question, asking about the gender of the talker shown during the test phase. In Experiments 1b and 1c, we made minor changes to the wording of the exit survey and removed some questions that had been found to be uninformative (for details, see SI).

Following the exit survey, a final 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.

Exclusions. We removed participants who (1) experienced technical difficulties or did not complete the experiment, (2) reported to not have used headphones or otherwise did not follow instruction, (3) did not answer the catch question about the talker's gender correctly, (4) had unusually fast or slow reaction times (participant's mean log-transformed RT outside of 3 SD of mean of participant means), or (5) had swapped the response keys, as determined by their responses. For this purpose, we considered participants with significant slopes in the opposite of the expected

direction as likely having swapped the response keys.[4]

Table 1 summarizes the participant exclusions for all experiments. After participant exclusions, we applied trial-level exclusions. For Experiments 1a-c, we excluded 84 trials (0.7%) that were missing a categorization response since participants indicated a catch trial (as detailed under *Procedure*, the experiments reported here did not actually contain any catch trials). We also excluded 213 trials (1.7%) with unusually fast or slow reaction times (within-participant scaled log-transformed RTs outside of 3 SD of mean of scaled log-transformed RTs at that trial position; for details and visualization, see SI). This left 12,472 observations from 177 participants across the three experiments. Finally, if trial-level exclusions resulted in more than 10% missing responses, participants were also excluded.

Table 1 Participant exclusions for all experiments reported. Total exclusions can be less than the sum of all individual exclusion criteria since some participants failed multiple criteria. The SI contains additional visualizations of participant exclusions.

Experiment	**1a**	**1b**	**1c**	**2**	**2b**
Recruited	*64*	*64*	*64*	*64*	*64*
Technical difficulty	-	-	2 (3.1%)	-	-
Did not follow instructions	-	3 (4.7%)	1 (1.6%)	1 (1.6%)	14 (21.9%)
Swapped keys	-	1 (1.6%)	1 (1.6%)	2 (3.1%)	-
Catch question	-	-	-	-	-
Outlier RT	3 (4.7%)	1 (1.6%)	-	2 (3.1%)	2 (3.1%)
Too many missing trials	1 (1.6%)	1 (1.6%)	1 (1.6%)	1 (1.6%)	-
Total	*4 (6.3%)*	*6 (9.4%)*	*5 (7.8%)*	*6 (9.4%)*	*16 (25%)*

[4] This differs somewhat from the approach taken in Liu and Jaeger (2018), who included only participants whose categorization functions had significant effects of the *asi-ashi* continuum in the expected direction. We instead also included participants whose categorization functions were 'flat' over the test continuum (no significant effect), excluding only participants with significant effects of the *asi-ashi* continuum in the opposite of the expected direction. We decided on this change prior to analysis, and consider it more adequate since participants with 'flat' categorization functions over our continuum are not *necessarily* uncooperative or misunderstanding the task (recall that the *asi-ashi* continuum we used did not include steps that would be expected to be perceived as either 100% *asi* or 100% *ashi*).

Results

Statistical power. No power analyses were conducted because the information gain would have been minimal: 1) other than the fact that we used audiovisual rather than audio-only stimuli, previous work has reliably detected effects of moderate size with the stimuli and design used here, including in the same web-based paradigm we employed here and even with fewer participants (e.g., 40 instead of 64 participants in Liu & Jaeger, 2019), 2) power simulations for those previous experiments found power >95% for moderate effect sizes even under conservative simulations with inflated inter-subject variability (ibid); and 3) we planned multiple replications of the critical test (Experiments 1a-c).

Analysis approach. We use Bayesian generalized linear mixed-effects 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). Responses (1 = *ashi* vs. 0 = *asi*) were regressed against pen location (effect coded: .5 = in mouth vs. -.5 = in hand), visual bias (effect coded: .5 = /ʃ/-bias vs. -.5 = /s/-bias), acoustic continuum, and test block as well as all their interactions. The six continuum steps and the six test blocks were coded as monotonically ordered categorical predictors (Bürkner & Charpentier, 2020). This avoids the linearity assumption made in most previous analyses of perceptual recalibration experiments, allowing changes across continuum steps or from block to block to have non-linear effects, while still constraining effects to be monotonic.[5]

All analyses further contained the full random effect structure for the three design variables pen location, visual bias, and acoustic continuum (by-participant intercepts and slopes for all

[5] Block was included in the analysis to provide a baseline for planned subsequent experiments on question (2) mentioned in the Open Science Statement. The inclusion does not, however, change any of the results. Additional analyses strongly supported linear effects of acoustic continuum and non-linear effects of test blocks for all experiments. The results we report below replicate when standard linear effects are used for continuum and block.

population-level predictors). No random slopes for test block were included since our studies were not designed to test this nuisance effect, leading to convergence problems for some experiments.

We followed recommended practice and used weakly regularizing priors to facilitate model convergence —specifically, the exact same practice as in our previous work to reduce researchers' degrees of freedom (e.g., Hörberg & Jaeger, 2021; Xie, Liu, & Jaeger, 2021). For fixed effect parameters, we used Student priors centered around zero with a scale of 2.5 units (following Gelman et al., 2008) and 3 degrees of freedom. For the monotonic predictors, we used a Dirichlet prior with the default $\alpha_1 = \cdots = \alpha_j = 1$. For random effect standard deviations, we used a Cauchy prior with location 0 and scale 2, and for random effect correlations, we used 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 $\hat{R} \leq 1.002$). All analyses were fit using the library *brms* (Bürkner, 2017) in R version 4.3.2 (R Core team, 2023).

Hypothesis tests. The SI lists the full model summary for all analyses. In the main text, we present Bayesian hypothesis tests over the fitted GLMMs for the questions of interest. Additionally, we report whenever the bidirectional 95% credible interval for any other effects does not contain 0. This was not the case for any effects in Experiments 1a-c. Table 2 summarizes those tests for all three experiments. Here and for all other experiments, the effects of all other predictors were assessed for the *first* test block and while marginalizing over continuum steps (following Liu & Jaeger, 2018, 2019). Figure 2 shows participants responses depending on the pen location, acoustic continuum and visual bias.

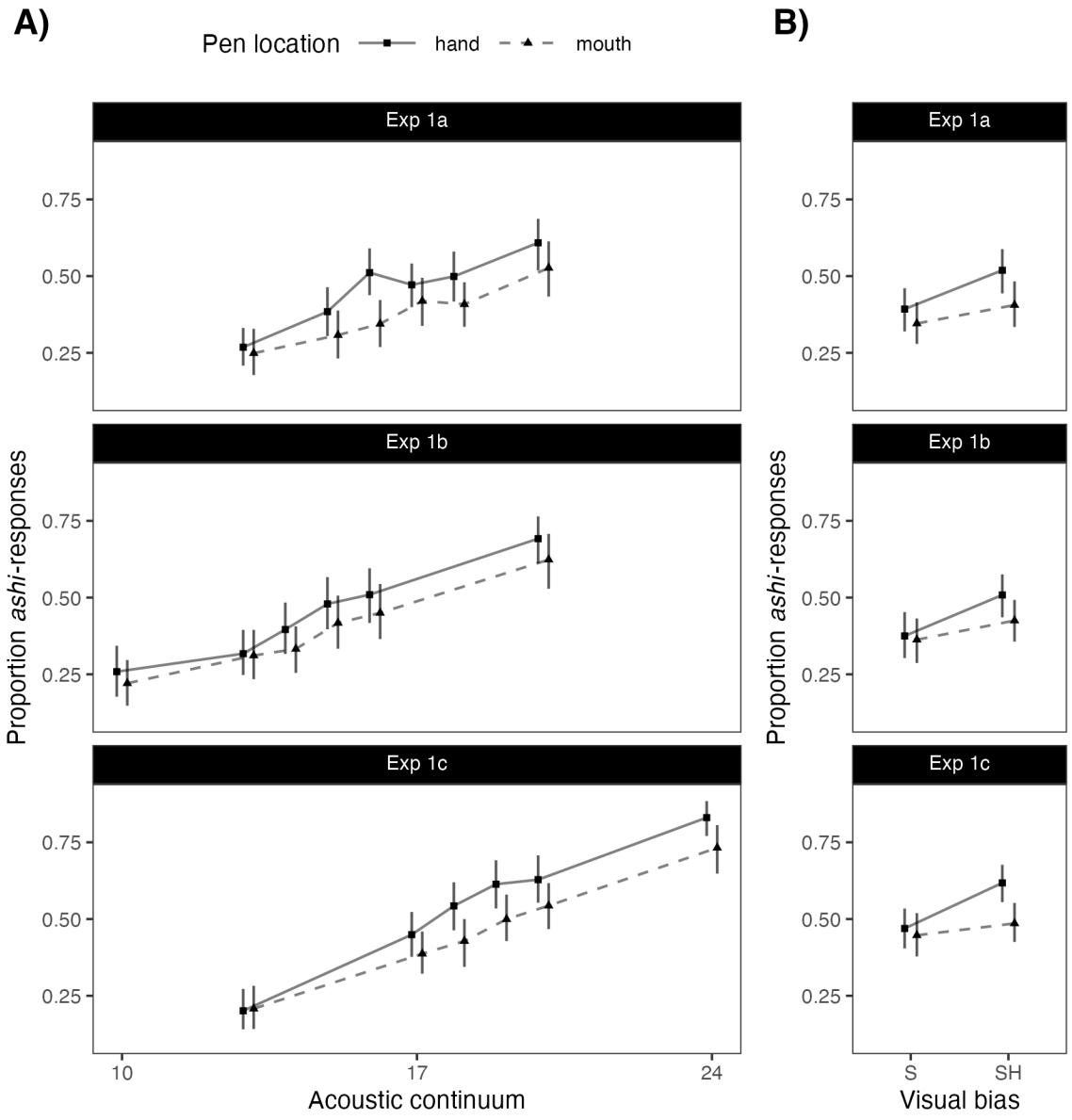

Figure 2 Summary of participants' responses in Experiments 1a-c, depending on pen location and acoustic continuum step (Panel A) or visual bias (Panel B). Points show means of by-participant averages. Intervals show bootstrapped 95% CIs over those by-participant means. Labels along the X-axis numbers refer to the 31-step continuum created by Liu & Jaeger (2018), where 1 and 31 were clear asi *and* ashi *endpoints, respectively.*

Table 2 *Summary of hypothesis tests based on GLMM analyses for Experiments 1a-c. Hypotheses about the effects of the pen are shown in the top four rows. Hypotheses about the effects of acoustic and visual biases are shown in the middle three rows. Hypotheses about how the effects (do not) change across blocks are shown in the bottom three rows. Hypotheses for which we had no strong expectations are shown with shaded backgrounds.*

	Exp 1a			**Exp 1b**			**Exp 1c**		
	$\hat{\beta}$	BF	$p_{posterior}$	$\hat{\beta}$	BF	$p_{posterior}$	$\hat{\beta}$	BF	$p_{posterior}$
Pen in mouth → fewer *ashi*-responses	-.75	284.7	.996	-.34	12.1	.924	-.47	16.9	.944
More *ashi*-biased *acoustically* → larger pen effect	-.03	1.8	.643	-.03	2.8	.740	-.08	30.0	.968
More *ashi*-biased *visually* → larger pen effect	-.50	11.0	.917	-.23	2.7	.731	-.90	63.5	.984
More *ashi*-biased acoustically & visually → even larger pen effect	-.08	3.2	.760	-.01	1.2	.536	-.05	2.6	.721
More *ashi*-biased *acoustically* → more *ashi*-responses	.20	>3999	>.999	.24	>3999	>.999	.29	>3999	>.999
More *ashi*-biased *visually* → more *ashi*-responses	.48	66.8	.985	1.05	>3999	>.999	.43	25.0	.962
Acoustic and visual effects are independent	-.12	10.9	.916	.01	52.5	.981	.02	49.6	.980
Pen effect stable	.10	25.2	.962	.02	134.3	.993	.08	54.4	.982
Acoustic *ashi*-bias stable	.03	2.6	.719	.02	64.2	.985	.02	3.1	.755
Visual *ashi*-bias effect stable	-.04	150.9	.992	-.13	18.1	.948	.05	87.7	.989

Of primary interest, participants in all three experiments were less likely to respond *ashi* if the pen was in the mouth (BFs > 12.1), as predicted by the compensation hypothesis. There also was evidence that this effect increased for stimuli that were acoustically or visually more *ashi*-like. This evidence was strongest for Experiment 1c (BFs > 30), potentially because the effect of compensation—a decrease in the probability of *ashi*-responses—is more difficult to detect for audiovisual stimuli for which *ashi*-responses are unlikely to being with. Similar trends were, however, present across all three experiments.

Beyond the effect of primary interest, all three experiments exhibited the expected effects of the acoustic continuum (BFs > 3999) and visual bias (BFs > 25), with increasing probabilities of *ashi*-responses when the audiovisual articulatory evidence biased towards *ashi*. These two effects seem to be independent of each other, suggesting additive effects of acoustic and visual evidence (BFs > 10.9, in line with models of ideal cue integration, see Massaro & Friedman, 1990; Bicknell, Bushong, Tanenhaus, & Jaeger, 2024). All three experiments also suggest that the effects of pen location, visual bias, and the acoustic continuum were generally stable across blocks (all BFs > 1), though the strength of the evidence in favor of this hypothesis varied between effects and experiments, and was merely anecdotal in some cases (1 < BFs < 3).

Finally, while not of particular relevance to our goals, we note that the choice of the acoustic continuum steps—which differed across experiments—clearly affected participants' perception (as also found in, e.g., Yamada & Tohkura, 1992). This is evident, for example, when one compares the proportion of *ashi*-responses for the lowest continuum step of Experiment 1c against the acoustically identical step in Experiments 1a and 1b in Figure 2.

Discussion

Experiment 1a-c tested whether presence of a pen in a talker's mouth affects listeners' perception of an audiovisual /s/-/ʃ/ continuum. All three experiments find this to be the case, despite

variation in the specific acoustic continuum steps employed by each experiment. Specifically, listeners were more likely to categorize an audiovisual input as *asi* when the talker in the video had a pen in the mouth, compared to when the talker held the pen in the hand. This effect was larger for tokens that were acoustically or visually more *ashi*-like, closely resembling findings for compensation for visually presented phonetic context (Kang et al., 2016).

Crucially, the directionality of our effects suggest compensation rather than ordinary cue integration. It is well established that non-phonetic, non-acoustic factors are integrated in perception. For example, Gick & Derrick (2009) found that feeling a burst of air on the skin—consistent with the aspiration of a /p/ but not with /b/—influenced listeners' perception of a VOT continuum, without conscious knowledge of the manipulation. However, the directionality of this effect was integratory rather than compensatory: the puff of air promoted increased /p/ responses. Our results, in contrast, are unexpected if listeners simply integrated visual and acoustic evidence of articulation, without discounting the *causes* for that evidence. The presence of a pen is expected to increase lip rounding and oral cavity opening. Either of these would result in lower center of gravity (similar to the effects of a bite-block, McFarland & Baum, 1995; Baum et al. 1996), making a sound acoustically more /ʃ/-like. If listeners naively integrated this visual evidence with the acoustic evidence, listeners should be *more* likely to respond *ashi* when the pen is in the mouth—the opposite of what we observed in all three experiments. Similarly, if listeners ignored the pen, or if the effects of the pen on articulation were not sufficiently visually evident, we should have failed to find *any* effect of pen location. This was not the case. Instead, the results of Experiments 1a-c are predicted by the hypothesis that listeners expect and 'explain away' the effect of the pen, paralleling compensation effects previously documented for surrounding phonetic context.

One alternative explanation would be that the pen partially or completely obscures some of the visual cues to /ʃ/—i.e., rather than causing more lip rounding or a more open oral cavity, the pen might obscure the presence of lip rounding and cause the oral cavity to be more closed. This

would explain the observed direction of the effect of pen location, and its enhancement for visually more *ashi*-like stimuli. It would, however, fail to predict why the effect of pen location increases for acoustically more *ashi*-like stimuli. Nevertheless, Experiment 2 further addresses this possibility.

Experiment 2

The materials and procedure of Experiment 2 were identical to Experiment 1c, except that the talker's mouth was occluded by a black rectangle during the production of the /s/-/ʃ/ fricative (see Figure 3). The rectangle was absent at the start and end of the video, appearing at the start of the fricative and disappearing at the end of the fricative. This left it very apparent *that* the pen was in the mouth during the production of the fricative, while occluding most direct evidence of the effect of the pen on the specific state of the articulators (lip rounding, oral cavity opening) during the production of the fricative (see Viswanathan & Stephens, 2016 for a similar occluder manipulation to remove visual *articulatory* context).

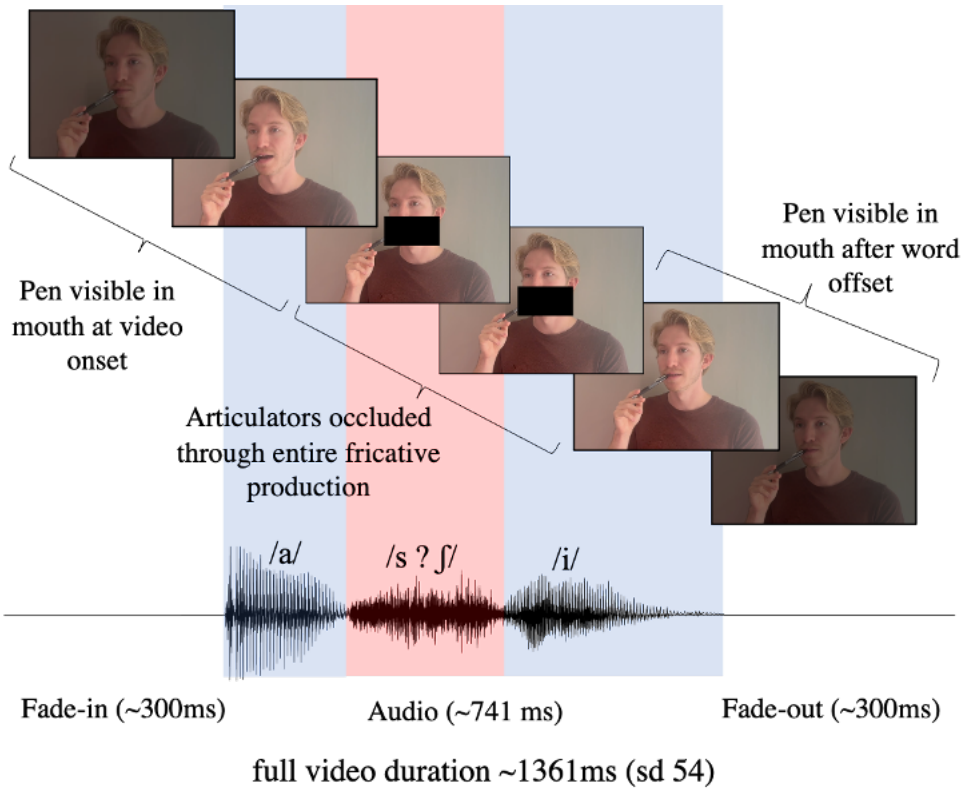

Figure 3 *Illustrating the critical manipulation of Experiment 2. As in Experiments 1a-c, the talker either had the pen in the mouth (left) or in the hand (right) during the production of the fricative. Unlike in Experiments 1a-c, a black box occluded the talker's mouth during the production of the fricative.* ***Note that actual stimuli in Experiment 1a-c (available at osf.io/2asgw) used a female talker.***

Experiment 2 served two purposes. First, by assessing the effect of pen location in Experiment 2, we can test whether the presence of a pen in the mouth was sufficient to cause the effect observed in Experiments 1a-c or whether listeners need to have more direct evidence of the *articulatory consequences* of the pen in the mouth. For example, if listeners only compensate if they observe that the pen indeed causes more lip rounding or larger opening of the oral cavity during the production of the fricative, then we expect the effect of the pen—replicated three times in Experiments 1a-c—to be no longer observed in Experiment 2. This latter possibility is predicted by compensation accounts like that advanced by Fowler since "it does not matter why the lips were rounded; it only matters that they were rounded" (Fowler, 2006, p. 166).

Evidence from visually presented *phonetic* context seem to be compatible with Fowler's conjecture. Previous work has found effects of visually presented phonetic context to be strongest when the relevant visual evidence—e.g., of lip-rounding—is particularly clear (e.g., Mitterer, 2006; Kang et al., 2016 vs. Vroomen & de Gelder, 2010) and when it is still present during the articulation of the target sound on which compensation is assessed (Holt, Stephens, & Lotto, 2005; for discussion, see Fowler, 2006; Lotto & Holt, 2006). Experiment 2 tests whether the same holds for the effects of the pen that we observed in Experiments 1a-c.

Second, Experiment 2 allows us to test whether the decreased rate of *ashi*-responses when the pen was in the mouth in Experiments 1a-c was due to visual occlusion of articulatory evidence, rather than compensation. Under this alternative hypothesis, both pen conditions (pen in mouth vs. hand) of Experiment 2 should yield rates of *ashi*-responses comparable to the pen in mouth condition in Experiment 1c (since Experiment 2 occludes most direct visual evidence of fricative articulation). Thus, Experiment 2 aimed to distinguish three hypotheses, two of which are elaborations of the compensation hypothesis: (1a) that listeners compensate for the visually evident presence of a cause that is known to affect the production of the fricative (pen in the mouth), (1b) that listeners compensate based on the visually evident state of the articulators caused by the pen in the mouth, rather than the presence of the pen itself, (2) that the effects of Experiments 1a-c were due to visual occlusion of articulatory cues, rather than compensation.

Methods

Participants. We again recruited 64 participants, using the same approach, payment, etc. as in Experiment 1c. Participants took an average of 22.3 minutes to complete the experiment (SD = 17.3 minutes).

Materials. All materials were the same as in Experiment 1c, except for the addition of a black rectangle to the video, as described above (Figure 3). The black rectangle was positioned

such that vertically, the area from the bottom of the talker's nose to the bottom of her chin were blocked from view. Horizontally, the entire width of the face was occluded. This was intended to occlude visually specified articulation, including lip rounding, mouth aperture, and tongue position. In cases where the talker moved during production, the size of the rectangle was increased such that the above criterion always applied. This gave rise to slightly different dimensions between different video frames. The occluder appeared during the video frame after the talker's maximum mouth aperture for the preceding vowel. The occluder disappeared at word offset. The vowel after the fricative was therefore also visually occluded. This window was intended to balance the competing constraints of giving subjects maximum opportunity to see the pen in the talker's mouth, while blocking the entirety of the fricative segment.

Procedure. The procedure was identical to Experiment 1c, with the exception that the phrase "with a black box occluding the speaker's mouth" was added to instructions where relevant**.**

Exclusions. We applied the same exclusion criteria as in Experiments 1a-c, removing six participants (9.4%; see Table 1). After participant exclusions, 11 trials (0.3%) were missing observations due to (incorrect) catch trial responses and an additional 97 trials were excluded for irregular RTs (2.3%), leaving for analysis 4069 observations from 58 participants.

Results

We used the exact same analysis approach as in Experiments 1a-c. The SI lists the full model summary for all analyses. Table 3 summarizes the hypothesis tests, Figure 4 shows participants' responses with those from Experiment 1c shown in the background for comparison. In contrast to Experiments 1a-c, we found no evidence for a main effect of pen location (BF = 1.0), Similarly, the effect of visual bias on participants' responses was also substantially reduced, though still in the same direction as in Experiments 1a-c (BF = 1.8). Participants continued to be strongly affected by the acoustic continuum (BF > 3999), the effect of which was similar and numerically

greater in magnitude ($\hat{\beta} = .31, SE = .028$) to Experiments 1a-c ($\hat{\beta}s$ between .20 and .29).

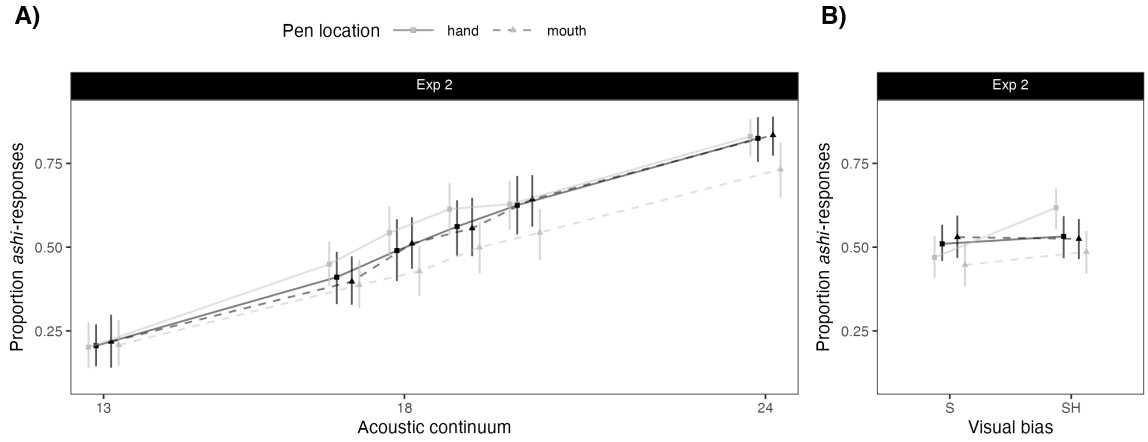

Figure 4 Summary of participants' responses in Experiments 2, depending on pen location and acoustic continuum step (Panel A) or visual bias (Panel B). For comparison, the results from Experiment 1c are shown in the background. The two experiments were identical except for the presence of the black rectangle during the production of the fricative in Experiment 2

Table 3 Summary of hypothesis tests based on GLMM analyses for Experiment 2. Hypotheses about the effects of the pen are shown in the top four rows. Hypotheses about the effects of acoustic and visual biases are shown in the middle three rows. Hypotheses about how the effects (do not) change across blocks are shown in the bottom three rows. Hypotheses for which we had not strong expectations are shown with shaded backgrounds.

	Exp 2		
	$\hat{\beta}$	BF	$p_{posterior}$
Pen in mouth → fewer *ashi*-responses	.00	1.0	.501
More *ashi*-biased *acoustically* → larger pen effect	.01	.7	.407
More *ashi*-biased *visually* → larger pen effect	-.17	2.1	.675
More *ashi*-biased acoustically & visually → even larger pen effect	-.05	1.9	.658
More *ashi*-biased *acoustically* → more *ashi*-responses	.31	>3999	>.999
More *ashi*-biased *visually* → more *ashi*-responses	.07	1.8	.648
Acoustic and visual effects are independent	-.04	45.1	.978
Pen effect stable	.01	132.6	.993
Acoustic *ashi*-bias stable	.02	35.6	.973
Visual *ashi*-bias effect stable	-.01	144.0	.993

Discussion

These results suggest that participants in Experiment 2 paid attention to the stimuli, and yet failed to exhibit any effects of pen location. In the SI, we report Auxiliary Experiment 2b. This experiment was identical to Experiment 2, except that participants had to *additionally* press the SPACE bar whenever the pen was in the talker's mouth. This was intended to (and did successfully) direct participants' attention towards the location of the pen. Experiment 2b replicated all effects of Experiment 2—including the absence of a credible effect of pen location ($\hat{\beta}$= .17, BF = .4, $p_{posterior}$ = .26).[6] Together, the findings of Experiments 2 and 2b thus suggest that a pen in the mouth of the talker is *not* sufficient to elicit the effect observed in Experiments 1a-c.

The comparison of Experiment 2 against Experiment 1c in Figure 4A further suggests that the effects of pen location in Experiments 1a-c are unlikely to be exclusively due to the pen

[6] Participants did, however, struggle with the more complex task of Experiment 2b, leading to a higher rate of participant exclusions (> 25%).

occluding visual cues to /s/-/ʃ/: at least for the two most *ashi*-like audio stimuli (for which the effect of pen location was strongest in Experiments 1a-c), responses in Experiment 2 seem to group with the pen-in-hand (no occlusion), rather than pen-in-mouth, condition in Experiment 1c. There is, however, also some evidence that visual occlusion might explain *part* of the effects in Experiments 1a-c. For the two steps in the middle of the acoustic continuum, responses in Experiment 2 fall half-way between the pen-in-hand and pen-in-mouth conditions of Experiment 1c (for the remaining two steps, the effect of pen location was too small even in Experiment 1c to draw meaningful conclusions about Experiment 2). Additional analyses presented in the SI confirmed that the pen-in-mouth condition in Experiment 1c resulted in fewer *ashi*-responses than the visual occluder in Experiment 2 ($\hat{\beta}$= -.66, BF = 38.6, $p_{posterior}$ = .975), whereas the visual occlusion in Experiment 2 did *not* result in in fewer *ashi*-responses than the pen-in-hand condition in Experiment 1c ($\hat{\beta}$= .12, BF = 0.6, $p_{posterior}$ = .361). Indeed, the only striking similarity between the pen-in-the-mouth and visual occlusion was that both reduced the effect of visual bias (see SI for details, and also Figure 4B). This is expected given that those visual cues were masked by the black rectangle in Experiment 2.

At first blush, the results of Experiment 2 favor hypothesis (1b) described in the introduction to Experiment 2: that listeners compensate based on the visually evident state of the articulators caused by the pen in the mouth rather than the presence of the pen itself. It is, however, possible that the addition of the black box had effects beyond removing visual evidence about the state of the articulators. Here we briefly discuss two possibilities that might require attention in future research.

First, the black occluder box might distract participants from the primary task, which could interrupt the cognitive process underlying compensation. While we cannot conclusively rule out this possibility, we consider it unlikely for a number of reasons. Compared to Experiments 1a-c, listeners in Experiment 2 demonstrated about the same—in fact, numerically somewhat greater—

sensitivity to the acoustic continuum (see Figure 4A), speaking against a higher proportion of attentional lapses or random guesses. Along a similar vein, average reaction times were not significantly higher in Experiment 2 (mean log-RT = 3.44, SD = 0.15), compared to Experiments 1a-c (mean = 3.45, SD = 0.14). Finally, while the black box might have been initially surprising to listeners, it appeared very predictably at approximately the same point in each of the 72 video trials. Thus, any initial distraction would likely atrophy over the course of testing. Analyses reported in the SI confirm that no effects emerged or interacted with testing block.

A second alternative is that while the presence of the pen itself—rather than its impact on the articulators—may be sufficient to induce compensation, the black box made the location of the pen less obvious and harder to visually access. The exit survey ameliorates this concern: the proportion of participants who mentioned the pen in the free-response post-experiment survey did not significantly decrease in Experiment 2 (29.7%) compared to Experiment 1a-c (36.7%; for statistical tests and visualizations, see SI). Additionally, Experiment 2b—which explicitly asked participants to press SPACE whenever the pen was in the talker's mouth—also failed to find the compensation effects we obtained in Experiments 1a-c. However, the dual-task nature of Experiment 2b imposed additional and orthogonal attentional demands (for details, we refer to the SI).

Future work could more conclusively assess the role of the pen versus its impact on the articulators by removing visual information more selectively. For example, videos could be presented of pen-impacted articulation, but with the pen itself removed in video post-processing. Conversely, video editing could be used to create an occluder that hides visual information about the articulators and face, while still showing the pen and hand of the talker. Either of these manipulations would require substantially more advanced video-editing than used in Experiment 2.

General Discussion

Taken together the results of Experiments 1a-c and 2, suggest that listeners compensate for visually evident effects of the pen on the configuration of articulators that are relevant to the /s/-/ʃ/ contrast. This suggests that speech perception can normalize or 'explain away' at least some effects of the pen on articulation, if they are sufficiently visually evident. To the best of our knowledge, this is the first demonstration that non-phonetic, visually evident effects on articulators—the pen's effect on lip shape—affect listeners' perception in ways consistent with compensation accounts. This provides novel support for Fowler's conjecture (Fowler, 2006): it indeed does not appear to matter why *exactly* a talker's lips exhibit a certain shape, as long as (1) they do, and (2) plausibly do so for some reason other than the articulation of the current segment (here, the pen in the mouth).

The results of Experiment 2—the absence of compensation in the absence of direct visual evidence that lip shape was indeed affected by the pen—replicates for non-phonetic context, what has previously been demonstrated for phonetic context: compensation for visually presented context seems to be substantially reduced or no longer observed when the relevant articulatory effects are not visually evident during the articulation of the target sound (Holt et al., 2005; Viswanathan & Stephens, 2016).

The present results also raise new questions for future research on compensation and adaptive speech perception more broadly. We briefly discuss two. First, our findings leave open whether compensation for visually evident effects of non-phonological causes—like the pen—draws on the same neural mechanisms as normalization/compensation for the effects of phonetic contexts. The present findings only suggest that Fowler's compensation account provides a unifying explanation for the qualitative consequences of both phenomena. It is unclear, for example, whether compensation for visually presented non-phonetic context takes place in the same brain areas that are responsible for audiovisual integration during speech perception, and whether those areas also process compensation for preceding phonetic context. Alternatively, compensation might

take place at multiple points in the processing of speech input.

Second and finally, the present results raise questions for future research on perceptual recalibration. As mentioned in the introduction, previous work has found that perceptual recalibration to an unfamiliar talker's speech can be blocked when the unexpected pronunciations occur while the talker has a pen in the mouth. In perceptual recalibration experiments, listeners are exposed to speech from an unfamiliar talker for which the realization of a particular sound is shifted towards a neighboring category. For example, Kraljic & Samuel (2006) exposed listeners to either typical /ʃ/ sounds and sounds ambiguous between /s/ and /ʃ/ but in lexical contexts favoring /s/ interpretation (e.g., *dinoshaur*, /s/-biased exposure) or to typical /s/ sounds and ambiguous sounds rather in /ʃ/-favoring contexts (e.g., *masinery*, /ʃ/-biased exposure), mixed with many filler stimuli. Following exposure, listeners were tested on an audio-only *asi-ashi* continuum. As is typical for such perceptual recalibration experiments, /ʃ/-biased exposure caused listeners to categorize more tokens along the test continuum as *ashi*, compared to /s/-biased exposure. In a thought-provoking follow-up, Kraljic et al. (2008) found that this perceptual recalibration effect—the difference between /ʃ/- and /s/-biased exposure during the audio-only test phase—is blocked when the talker had a pen in the mouth during the pronunciation of the critical shifted exposure tokens. When the pen was instead in the hand during the shifted tokens, listeners again exhibited perceptual recalibration.

While this blocking effect has since been replicated multiple times, the mechanisms underlying the effect remain unclear (see discussion in Kraljic & Samuel, 2011; Liu & Jaeger, 2018). For instance, Kraljic et al. (2008) attributed the blocking effect to "pragmatic" reasoning that prevents perceptual learning for "incidental" changes in pronunciation—like those arising a pen is in the mouth—while allowing perceptual learning for changes that are considered "characteristic" of the talker's speech. Based on a series of new experiments, Kraljic and Samuel (2011) revised this account, suggesting instead that audiovisual inputs with a pen in the mouth are

stored as separate exemplars that do not affect listeners' categorization decisions for audio-only speech inputs—i.e., the type of speech input that has been used in the test phase of *all* perceptual recalibration experiments on this question to date. However, based on additional experiments, Liu and Jaeger (2018) argued that the existing data were, in fact, more compatible with an account along the lines of the original proposal by Kraljic et al. (2008)—as inference under uncertainty about the *causes* for the unexpected pronunciations. Despite important similarities between these different accounts, all previous explanations of the blocking effect share that they appeal to *learning* and/or *memory*.

The results of our experiments raise the possibility of another, qualitatively different, explanation: participants might *compensate* for the pen in the mouth during exposure. This would place the explanation in *perception*, rather than learning or memory. Compensation would be expected to make the shifted /s/ tokens sound *less* shifted (as it should make them sound less /ʃ/-like) and to make the shifted /ʃ/ tokens sound *more* shifted (as it makes them sound more /s/ like). If such compensation takes place before perceptual recalibration—or, put differently, if recalibration operates over compensated percepts—this would be expected to affect the outcome of recalibration. Without further considerations, compensation should weaken the effect of the /s/-biased exposure and strengthen the effect of /ʃ/-biased exposure, leaving it unclear how these two effects trade off. However, it is also known that shifts larger than those typically used in perceptual recalibration experiments *reduce* the effectiveness of exposure, because they reduce the rate at which participants still accept the shifted recordings as an instance of the intended sound (Babel et al., 2019). Babel and colleagues refer to this as the "goldilocks zone": perceptual recalibration is most effective when the sounds are shifted as far as possible towards the other sound *while still being perceived as an instance of the intended sound*. Compensation thus might indeed offer a particularly parsimonious explanation of blocked perceptual recalibration: the pen in the mouth during exposure reduces the effect of /s/-biased exposure because it makes the critical recordings

sound less shifted, and it reduces the effect of /ʃ/-biased exposure because it makes the critical recordings sound shifted *too far* to still be accepted as /ʃ/. If an explanation along these lines turns out to be correct, this would preempt the need to evoke memory (Kraljic & Samuel, 2011) or learning mechanisms (Kraljic et al., 2008). We consider this an interesting possibility to be explored in future research.

Acknowledgements: This research was supported by a Bilski-Mayer Summer Research Fellowship to SNC and by the University of Rochester through funding for *BCS 206: Undergraduate Research in Cognitive Science*.

Earlier versions of this work were presented at the University of Rochester Undergraduate Research Expo, and the 62nd annual meeting of the Psychonomic Society. We owe many thanks to other students and instructors of BCS 206/7 (class of 2020-2021), who provided invaluable feedback throughout the project. We thank Effie Kapnoula, Rachel Sabatello, and two anonymous reviewers for constructive feedback on a previous version of this manuscript. We also thank Carol Fowler, Arty Samuel, and Jean Vroomen for helpful pointers to relevant literature, and helping us better understand the relevant theoretical space. We gratefully acknowledge that this project is made possible through other researchers' dedication to open science: the audio recordings were obtained from Tanya Kraljic and Arthur Samuel, the video recordings were obtained from Molly Babel; web experimentation through Proliferate developed by Sebastian Schuster (ALPs Lab, Stanford); the JavaScript, HTML, and CSS for the experiment is based on open source code originally developed by Dave Kleinschmidt (Human Language Processing Lab, University of Rochester, https://github.com/hlplab/JSEXP). We thank Zach Burchill, Wednesday Bushong, Linda Liu, and Xin Xie for sharing materials and feedback for a tutorial on crowdsourcing experiments developed as part of this project (https://github.com/hlplab/Tutorial-MTurk-experiments-via-mturkutils).

Author contributions: SNC proposed the experiment. SNC, GEK, 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 GEK programmed the web-based experiments. GEK 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.

References

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

Baum, S. R., McFarland, D. H., & Diab, M. (1996). Compensation to articulatory perturbation: Perceptual Data. *The Journal of the Acoustical Society of America, 99(6),* 3791–3794.

Bejjanki, V.R., Clayards M., Knill D.C., Aslin R.N. (2011). Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception. *PLoS ONE, 6(5).*

Bürkner, P. C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*, 1–28.

Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420–451.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics, 38(2),* 167–184.

Cummings, S., Karboga, G. E., Yang, M., & Jaeger, T. F. (2025, January 21). Cummings, Karboga, Yang, & Jaeger. Compensation in audiovisual speech perception: discounting the pen in the mouth. Retrieved from osf.io/2asgw

Cummings, S. N., & Theodore, R. M. (2023). Hearing is believing: Lexically guided perceptual learning is graded to reflect the quantity of evidence in speech input. *Cognition*, *235*, 105404.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance, 26(3)*, 877–888.

Fowler, C. A. (2004). Speech as a Supramodal or Amodal Phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 189–201). Boston Review. https://doi.org/10.7551/mitpress/3422.003.0016

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics, 68(2),* 161–177.

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, *119*(3), 1712–1726.

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

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2(4), 1360–1383.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*(7272), 502-504.

Holt, L. L., Stephens, J. D., & Lotto, A. J. (2005). A critical evaluation of visually moderated phonetic context effects. *Perception & Psychophysics*, *67*, 1102–1112.

Hörberg, T., & Jaeger, T. F. (2021). A rational model of incremental argument interpretation: The comprehension of Swedish transitive clauses. *Frontiers in Psychology, 12*.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America, 125(6)*, 3983–3994.

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

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for

Coarticulation. *Speech Communication, 77*, 84–100.
Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*, 262–268.
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.
Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100(9),* 1989–2001.
Lindblom, B. E., & Sundberg, J. E. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. The Journal of the Acoustical Society of America, 50(4B), 1166-1179.
Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174* (June 2017), 55–70.
Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *45*(12), 1562–1588.
Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*, 178–183.
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
Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics, 28(3)*, 213–228.
Mann, V., & Soli, S. D. (1991). Perceptual order and the effect of vocalic context on fricative perception. *Perception & Psychophysics, 49(5),* 399–411.
McFarland, D. H., & Baum, S. R. (1995). Incomplete compensation to articulatory perturbation. *The Journal of the Acoustical Society of America, 97,* 1865–1873.
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
McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118(2),* 219–246.
McMurray, B., & Jongman, A. (2016). What comes after /f/? Prediction in speech derives from data-explanatory processes. *Psychological Science, 27(1)*, 43–52.
McMurray, B., Rhone A., & Galle M. (2012). *FricativeMakerPro*
Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics, 68(7)*, 1227–1240.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204-238.
Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA.
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).
R Core Team (2023). _R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rosenblum, L. D., Dorsi, J., & Dias, J. W. (2016). The impact and status of Carol Fowler's supramodal theory of multisensory speech perception. *Ecological Psychology*, *28*(4), 262-294.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America, 70(4),* 976–984.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America, 79(4)*, 1086–1100.

Viswanathan, N., & Stephens, J. D. (2016). Compensation for visually specified coarticulation in liquid–stop contexts. *Attention, Perception, & Psychophysics*, *78*, 2341–2347.

Vroomen, J., & de Gelder, B. (2001). Lipreading and the compensation for coarticulation mechanism. *Language and Cognitive Processes*, *16*(5–6), 661–672.

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

Yamada, R. A., & Tohkura, Y. I. (1992). The effects of experimental variables on the perception of American English/r/and/l/by Japanese listeners. *Perception & psychophysics*, *52*, 376-392.

Yeni-Komshian, G. H., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America, 70(4)*, 966–975.