2025. Proc Ling Soc Amer 11(1). 6033. https://doi.org/10.3765/plsa.v11i1.6033.

The role of sentence context in perceptual learning of speech

Wednesday Bushong, Audrey Jia-Ling Chang, Molly E. Doherty, & Lee O. Mercado*

Abstract. A key issue in speech perception is its *lack of invariance*: mappings between acoustic features and phonemic categories are not one-to-one. Listeners can address this problem by adapting their perception to a talker's idiosyncratic accent. Previous studies have used lexical cues to examine adaptation, but in everyday speech talkers use full sentences. We investigated whether semantic context in full sentences occurring either before or after an acoustically-manipulated target word can trigger adaptation. We found that sentence context can be used for adaptation, but findings are mixed on whether the timing of context impacts learning.

Keywords. psycholinguistics; speech perception; perceptual recalibration; perceptual learning

1. Introduction. The core problem of speech perception is its *lack of invariance*: there is no one-to-one mapping from acoustic cue values to phonemes. Rather, there are many acoustic cues that are each partially informative (e.g., voicing is acoustically cued by voice-onset time, fundamental frequency, vowel duration, and other cues; Lisker & Abramson 1970). Despite this general lack of invariance, there is also striking systematicity: for example, there is much greater consistency within American English speakers raised in the Midwest compared to the community of all American English speakers; and there is even greater consistency within a single talker (for more discussion, see Kleinschmidt 2019). One way that listeners deal with the lack of invariance problem is to *adapt* their perceptual boundaries between speech categories for new talkers—a strategy that works precisely because talkers are self-consistent. One mechanism by which listeners can adapt is through supervised learning: if the listener is (reasonably) sure which speech sound a talker intended to produce, they can use that information to update their expectations about the talker's acoustic distribution for each phonemic category.

Perceptual recalibration is a paradigm designed to systematically test this type of supervised perceptual learning of speech sounds (Norris et al. 2003). Most of these studies employ a *lexically-guided* recalibration paradigm, which takes advantage of lexical cues to label words. Participants are exposed to target words that contain a sound contrast of interest—for example, /b/ vs. /p/, which contrast in voicing and can be acoustically manipulated along the voice onset time (VOT) continuum (VOT is the primary cue to stop voicing in American English; Lisker & Abramson 1970). Critically, these manipulated segments occur in words that do not have minimal pairs, like *bash* and *past*, biasing listeners to perceive the acoustically manipulated segment as the sound consistent with the lexical item (Ganong 1980). A VOT that was previously ambiguous can then be inserted into either the /b/-biasing or /p/-biasing word, pushing participants to perceive the ambiguous segment as more /b/-like or more /p/-like. Participants also hear clear examples of the contrastive sound paired with the appropriate lexical items. After exposure to many such items, participants then categorize a range of VOTs embedded in neutral contexts (like [?a]).

* Thank you to Dan LaMarche, Lay Espinal, and Zac Longo for assistance with stimulus development and manipulation, and to Sabrina Vaillancourt for lending her voice for recording. We would also like to thank Florian Jaeger and Xin Xie for helpful discussions about the perceptual recalibration paradigm. Authors: Wednesday Bushong, Wellesley College (wb104@wellesley.edu); Audrey Chang, Wellesley College (ac163@wellesley.edu); Molly Doherty, Wellesley College (mollydoh@icloud.com) & Lee Mercado, Wellesley College (lm122@wellesley.edu).

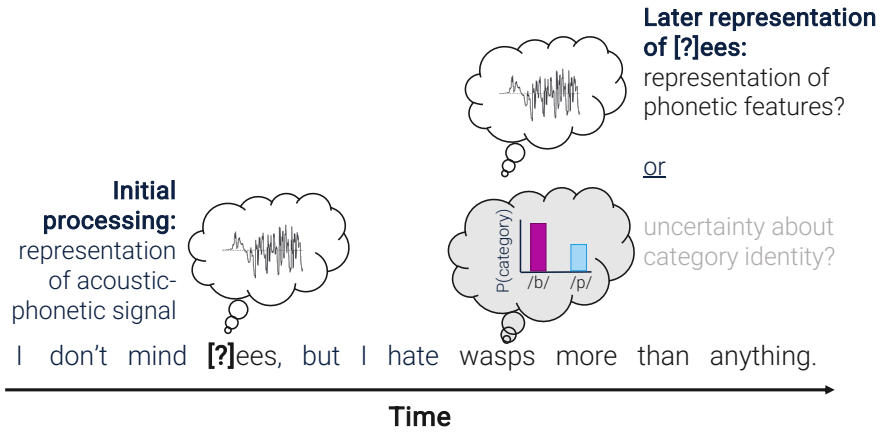


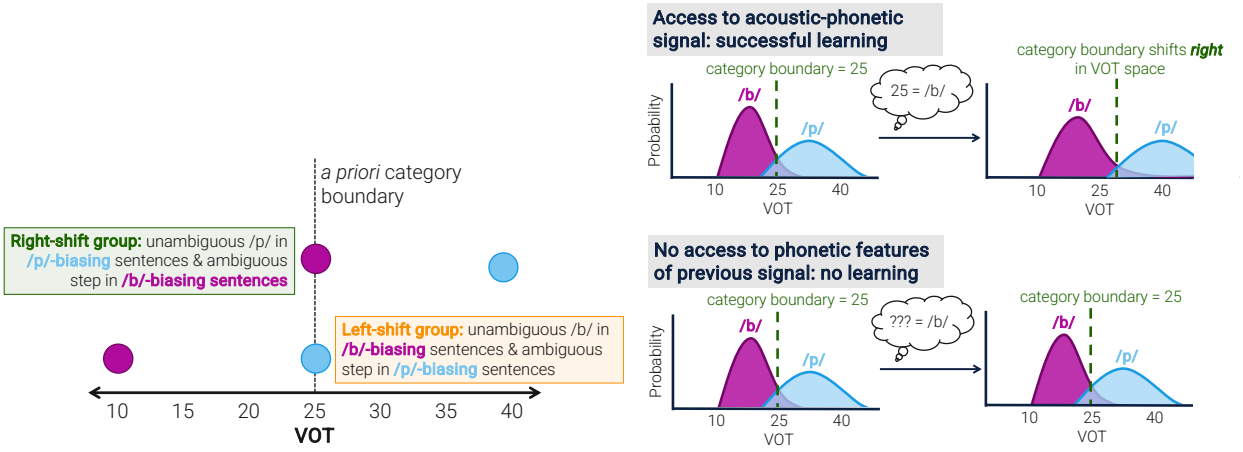
Figure 1. Possible representations of speech input over time.

If the participant has successfully learned the new acoustic distribution of /b/ and /p/, their categorization curves should correspond to their exposure: participants exposed to the ambiguous VOT in /p/-biasing words will have a lower VOT category boundary and vice versa for the other exposure group. These lexically-guided perceptual recalibration effects are strong—observed after only a few exposure trials (Vroomen et al. 2007)—and quite consistent across experiments.

An emerging literature has begun to investigate whether information occurring outside the critical target word can support adaptation. Jesse (2021) presented subjects with sentence contexts that were highly biased toward a particular final word, which contained an acoustically manipulated segment on an /s/-/f/ continuum, e.g., "Tom did not want the argument to turn into a real sight/fight". Unlike most perceptual recalibration studies, the critical words were minimal pairs, so participants could not rely on lexical information for disambiguation. Participants showed clear perceptual recalibration effects, suggesting that sentential context can aid in perceptual learning. These effects have been replicated with additional lexical manipulations and with other contrasts (Jesse 2024; Aoki & Zellou 2025). One notable exception is Luthra et al. (2021), who unlike the other experiments discussed here, used non-minimal pair critical target words, so that both the sentence context and the lexical bias of the target words provided information for learning. Listeners showed perceptual recalibration effects on the basis of the lexical context, but the addition of sentential context did not produce stronger effects.

This literature suggests that sentence context can help listeners recalibrate to manipulated sound contrasts. However, all of these studies used stimuli with sentence context that occurred *before* the critical target word of interest. In natural speech, it is common for relevant context to appear *after* a word of interest. Consider the example "I don't mind [?iz], but I hate wasps more than anything." The later word *wasps* suggests that the speaker's intended target was *bees* (as opposed to *peas*). While sentence context occurring before or after a critical target word can affect its perception (Connine 1987; Connine et al. 1991; Westwig et al. under review), it is unclear whether following sentence context can in turn aid in perceptual recalibration.

There are theoretical reasons to be interested in this question, because it can shed light on what level of representational detail listeners maintain about acoustic input over time. There is a rich debate about the timescale of representations during spoken language processing (for review, see Christiansen & Chater 2016). In particular, there is now very clear evidence that listeners



(a) (b)

Figure 2. Logic of perceptual recalibration experiments. Panel (a) shows the two major conditions in an experiment, right-shifted and left-shifted groups. Panel (b) shows how phonetic information is critical for learning: if listeners have access to phonetic detail when the sound is disambiguated to a particular category, they can use that knowledge to update their category beliefs, thus shifting their category boundary (top panel); if the listener no longer has access to the phonetic detail, they cannot appropriately update their beliefs about the category distributions.

can maintain some degree of subcategorical information about spoken input over time (McMurray et al. 2009; Brown-Schmidt & Toscano 2017; Bicknell et al. 2025: inter alia). Do listeners merely maintain uncertainty about category identity (Figure 1, bottom right), or do they maintain more fine-grained representations, such as acoustic details about the previous input (Figure 1, top right)? Perceptual recalibration provides a clear way to test this. To recalibrate requires that listeners alter their mapping of acoustic cues to categories; listeners must have access to both the acoustic value of the encountered input and its category identity at the same time in order to update this mapping (for further discussion, see Davis & Johnsrude 2007). If listeners can use information that occurs significantly later than the segment to perceptually recalibrate, then they must have maintained a representation about the relevant acoustic details over that period of time (in order for category identity information to arrive; see Figure 2).

There are two prior studies attempting to clarify the role of later context on perceptual learning of speech (Burchill et al. 2018; Caplan et al. 2021). These studies both expose participants to isolated words accompanied by subtitles varying in their timing relative to auditory presentation. Caplan et al. (2021) conducted a perceptual recalibration study where participants are exposed to single minimal-pair words (e.g., *[t/d]ent*) which are shifted to have shorter or longer VOTs than the average American English speaker. Target words were disambiguated by a subtitle that appeared either before or after the audio presentation. They found perceptual recalibration when subtitles were presented before the target word, but not after. In a similar vein but using a more naturalistic accent adaptation paradigm, Burchill et al. (2018) exposed participants to Spanish-accented English words which were subtitled either concurrent with, or after, audio presentation (at varying delays). In contrast with Caplan et al. (2021), they found that both concurrent and delayed subtitles aided accent adaptation above non-subtitled exposure.

Given the higher level of experimental control in Caplan et al. (2021) compared to Burchill et al. (2018), one might be tempted to conclude that post-target word context cannot aid in speech adaptation. However, there are a couple of reasons to think that the picture might be more complicated. First, Caplan et al. (2021) had an uncharacteristically long test phase compared to other perceptual recalibration experiments (162 trials, compared to 40-45 in Burchill et al. 2018; Jesse 2021, 2024; Aoki & Zellou 2025). Given that the acoustics of test tokens are uniformly distributed, listeners may continue to learn, effectively un-learning the exposure distribution (for further discussion, see Kleinschmidt & Jaeger 2015; Xie & Kurumada 2025). Second, and a more general concern with both subtitle studies, listeners' behavior differs when processing isolated words compared to words in a larger sentential context. For instance, studies testing for the effect of subsequent context on speech perception in isolated words find that contextual effects degrade before 250ms (Andruski et al. 1994); however, studies investigating similar effects in full sentences have found no decay in the effect of subsequent context at any time delay tested (up to 35 syllables after target word onset; Falandays et al. 2020; Bicknell et al. 2025). Thus, it is reasonable to hypothesize that subsequent sentence context may support perceptual recalibration in ways that subtitles of isolated words may not.

In the present study, we conduct three perceptual recalibration experiments where we expose participants to full sentences with contextual information occurring either before or after the word of interest.

2. Experiment 1.

2.1. Methods.

2.1.1. Ethical Approval.

All experiments were approved by the Brandeis Human Research Protection Program (Protocol 25109R-E).

2.1.2. Participants.

112 subjects were recruited via Prolific and were compensated $5.00 for their participation (based on a $10.00 per hour compensation rate). All participants were native speakers of American English currently living in the United States. We removed subjects whose accuracy on filler trials was below 80% (see below for details on filler trials). This resulted in the removal of 6 subjects from analysis, leaving 106 participants remaining whose data were analyzed.

2.1.3. Materials.

We used sentence materials which vary in whether supportive sentential context biases towards a /b/ or a /p/ interpretation of the target word, and whether context occurs before or after the target:

1(a) I don't mind **[?iz]**, but I hate **squash** more than anything. (p-biasing, context-after)

1(b) I don't mind **squash**, but I hate **[?iz]** more than anything. (p-biasing, context-before)

1(c) I don't mind **[?iz]**, but I hate **wasps** more than anything. (b-biasing, context-after)

1(d) I don't mind **wasps**, but I hate **[?iz]** more than anything. (b-biasing, context-before)

We used three different critical target word pairs across items: *bees/peas*, *bath/path*, and *beach/peach*. On average, biasing sentential context occurred a distance of 3-5 syllables from the target word (see Table 1).

Context and timing	Context bias distance (syllables)	Context bias distance (words)	Number of non-target words with b/p onset
b-biasing/before	4 (1.46)	3.84 (1.27)	0.42 (0.56)
b-biasing/after	4.19 (1.83)	3.84 (1.37)	0.48 (0.57)
p-biasing/before	3.65 (1.4)	3.48 (1.21)	0.45 (0.57)
p-biasing/after	4.26 (1.67)	3.9 (1.08)	0.55 (0.62)

Table 1. Statistics of sentence stimuli. Mean in main text of cell, standard deviation in parentheses.

We acoustically manipulated the voice onset time (VOT) of the first sound of the target word to vary between /b/ and /p/. VOT is one of the primary cues distinguishing voicing in American English stops, with shorter VOTs perceived as more /b/-like and longer VOTs as more /p/-like (Lisker & Abramson 1970). We followed the VOT manipulation procedure developed by Winn (2020). We confirmed successful acoustic manipulation in a norming study conducted when our lab developed these sentence materials (see Westwig et al. under review). In that norming study, we found the average category boundary was near 25ms VOT, with the floor of /p/-responses at 10ms, and the ceiling at 40ms; we thus chose the 25ms step as our 'ambiguous' token, 10ms as an 'unambiguous' /b/, and 40ms as an 'unambiguous' /p/. We used these steps to construct our recalibration conditions: in the left-shifted (or /b/-shifted) group, participants were exposed to the unambiguous /b/ token in /b/-biasing sentences, and the ambiguous token in /p/-biasing sentences. In the right-shifted (or /p/-shifted) group, participants were exposed to the ambiguous token in /b/-biasing sentences and the unambiguous /p/ token in /p/-biasing sentences.

In addition to our critical stimuli, we also presented participants with filler sentences where they made judgments about target words whose onset varied between /l/ and /r/. Like the critical stimuli, these sentences also contained context which biased toward one or the other interpretation; however, unlike the critical stimuli, these target words were not acoustically manipulated.

2.1.4. PROCEDURE. The experiment proceeded in two phases. In the exposure phase, participants listened to full sentences and were asked to categorize a target word in the sentence between two alternatives (e.g., for the example sentences in (1) above, participants are asked whether they heard "bees" or "peas" in the sentence). Participants heard 28 critical sentences in each context condition (/b/-biasing vs. /p/-biasing sentences), for a total of 56 critical exposure trials. Participants also heard 28 filler sentences containing /l/-/r/ contrasts. Stimulus order was fully randomized for each participant.

In the test phase, participants heard stimuli along a /b/-/p/ continuum in a neutral /a?a/ context. We used a seven-step VOT continuum from 10 to 40ms in 5ms steps. Each step was repeated 10 times for a total of 70 test trials. Stimulus order was fully randomized for each participant.

2.1.5. ANALYSIS. All analyses were conducted using R (R Core Team 2024). We fit a mixed-effects logistic regression model predicting /p/ responses from VOT (Gelman-scaled),[1] shift condition (sum-coded: -.5 = left-shifted condition, .5 = right-shifted condition), and context timing (sum-coded: -.5 = context-before, .5 = context-after); we also included the two-way interaction

[1] Gelman scaling is identical to z-scoring, but with two standard deviations used as the scaling factor (Gelman 2008).

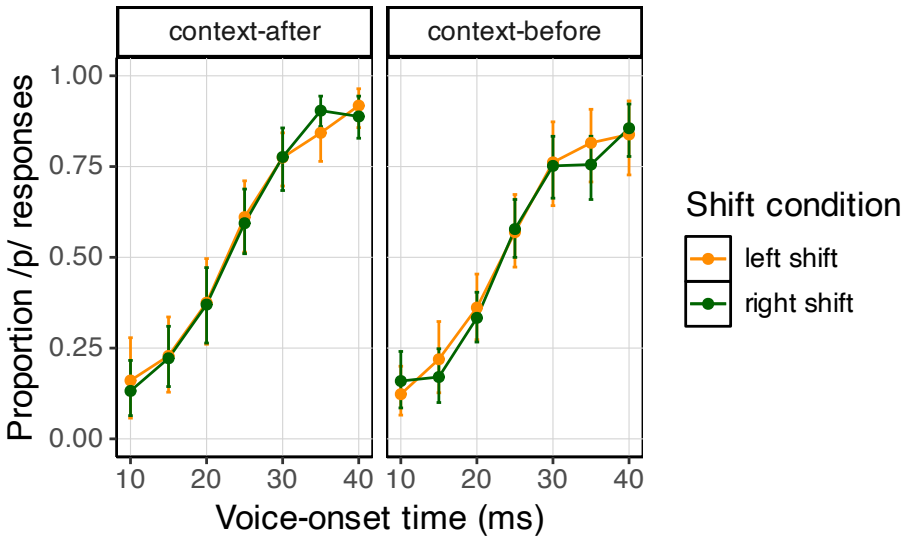


Figure 3. Experiment 1 /p/ responses (y-axis) during the test phase by VOT (x-axis), shift condition (color) and context timing (panel). Error bars are 95% confidence intervals bootstrapped over subject means.

between shift condition and context timing. We included the maximal random effects structure as specified by the design (Barr et al. 2013); in this case, because all manipulations were between subject, this corresponded to random intercepts by subject. We expect to observe an overall main effect of VOT, such that as VOT increases, participants' /p/ responses also increase. If participants successfully recalibrate, then there should be a significant main effect of shift condition such that the left-shifted group shows a higher proportion of /p/ responses than the right-shifted group. If participants learn equally well in both context timing groups, we should observe no interaction between shift condition and context timing; if, however, participants learn more effectively in one condition than the other we would expect to see such an interaction.

Even if there is a significant interaction between shift condition and context timing, it is possible that participants do show a learning effect in each condition that simply differs in magnitude. Thus, it is also worthwhile to conduct a simple effects analysis, which allows us to test whether there was a significant effect of shift condition *within* each context timing group. To that end, we fit a simple effects model, recoding context timing as a categorical variable and nesting shift condition.

In addition to our main analyses, we also analyze participants' responses in the exposure phase in order to ensure that our context manipulation was successful. We fit a mixed-effects logistic regression model predicting /p/ responses from context condition (whether the context biased toward a /p/ or /b/ interpretation), shift condition, and context timing condition, with random intercepts and random slopes by context bias by subject and by item. We expect to see a significant effect of context. There may also be a significant effect of shift condition with the right-shift condition showing higher /p/ responses, given the overall higher VOT values in the right-shift condition compared to the left-shift condition in exposure.

2.2. Results.

2.2.1. TEST PHASE. Figure 3 shows the empirical results of the Experiment 1 test phase. There was a significant main effect of VOT, such that /p/ responses increased with VOT value ($\hat{\beta} = 3.278, z = 42.255, p < .001$). There was no significant effect of shift condition ($p = .772$) or any other variables. In the simple effects model, there was no significant effect of shift condition in either context timing group ($ps > .671$).

2.2.2. EXPOSURE PHASE. In the exposure phase, there was a main effect of context such that /p/-biased context sentences led to higher /p/ responses ($\hat{\beta} = 6.066, z = 18.459, p < .001$). There was also a main effect of shift condition, such that the right-shifted group had higher /p/ responses ($\hat{\beta} = 2.05, z = 9.024, p < .001$). There were no other significant effects.

2.3. DISCUSSION. In Experiment 1, we conducted a perceptual recalibration study where we manipulated whether supportive sentential context occurred before or after a target word containing an acoustically manipulated segment. Surprisingly, we found no evidence for perceptual recalibration in either condition, despite the success of the context manipulation in the exposure phase and findings from previous work that sentence context aids perceptual recalibration (Jesse 2021, 2024; Aoki & Zellou 2025). However, there is a potential methodological concern: while all of the critical words in the exposure sentences contained the manipulated /b/-/p/ segment in word onset position, participants were tested on syllables where the /b/-/p/ segment appeared word medially. Given that the acoustics of stop voicing differ depending on their position within-word and within-syllable (Lisker & Abramson 1967), it is possible that our participants did not generalize their learning during the exposure phase to the tokens they were tested on.

In Experiment 2, we remedy this issue by testing participants on the target words they encountered in the exposure sentences. This ensures that participants are being tested on the same stimuli as they were exposed to, increasing our chances of finding a recalibration effect.

3. Experiment 2.

3.1. METHODS.

3.1.1. PARTICIPANTS. 161[2] participants were recruited via Prolific and were compensated $5.00 for their participation. All participants were native speakers of American English currently living in the United States who did not participate in Experiment 1. As with Experiment 1, we removed all participants with below 80% accuracy on filler trials. 15 participants were removed, leaving us with 146 participants whose data were analyzed.

3.1.2. MATERIALS. Sentence materials were identical to Experiment 1.

3.1.3. PROCEDURE. The procedures were identical to Experiment 1 during the exposure phase. In the test phase, participants categorized the /b/-/p/ target words from the exposure sentences presented in isolation (as opposed to the /a?a/ stimuli used in Experiment 1). The words were presented at the same seven VOT continuum steps used in Experiment 1 (10, 15, 20, 25, 30, 35, 40). Each target word was presented in each VOT step three times each, for a total of 63 test trials.

3.1.4. ANALYSES. Analyses were identical to Experiment 1 with one exception: since multiple target words were presented in test, we included random intercepts and VOT slopes by target

[2] We aimed to recruit 160 participants, but due to a technical error on Prolific, one extra participant was able to complete the study.

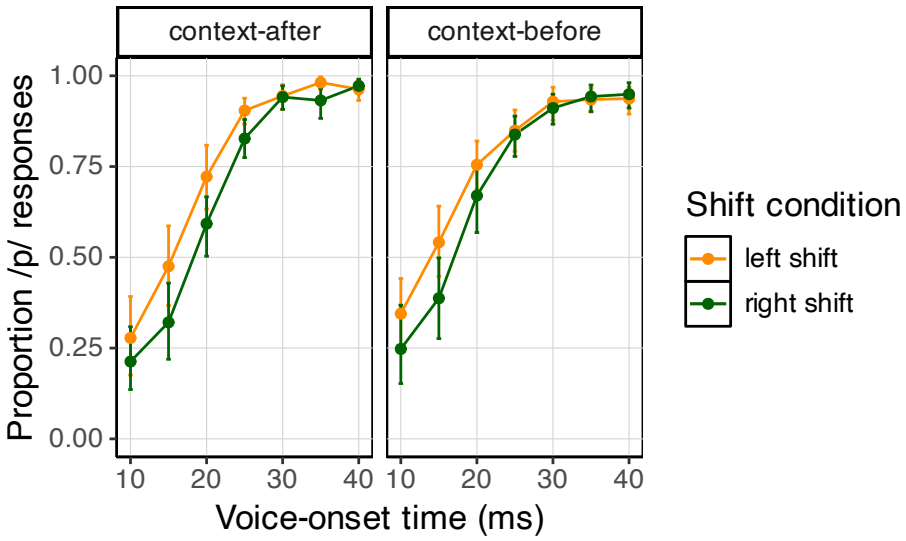


Figure 4. Experiment 2 /p/ responses (y-axis) during the test phase by VOT (x-axis), shift condition (color) and context timing (panel). Error bars are 95% confidence intervals bootstrapped over subject means.

word.

3.2. Results.

3.2.1. Test phase. Figure 4 shows the test phase results by shift and context timing condition. As in Experiment 1, we observed a significant effect of VOT on /p/-responses ($\hat{\beta} = 5.028, z = 13.084, p < .001$). Unlike Experiment 1, we did observe evidence for perceptual learning: there was a significant effect of shift condition on /p/-responses, such that right-shifted participants had a lower proportion of /p/-responses, consistent with a right-shifted category boundary ($\hat{\beta} = -.524, z = -2.253, p = .024$). There was not a significant interaction between shift condition and context timing ($p = .547$). The simple-effects model revealed that in the context-after group, there was a significant effect of shift condition ($\hat{\beta} = -.663, z = -2.003, p = .045$); but there was no significant effect of shift condition in the context-before group ($\hat{\beta} = -.383, z = -1.17, p = .241$).

3.2.2. Exposure phase. In the exposure phase, there was a main effect of context such that /p/-biased context sentences led to higher /p/ responses ($\hat{\beta} = 7.082, z = 22.235, p < .001$). There was also a main effect of shift condition, such that the right-shifted group had higher /p/ responses ($\hat{\beta} = 2.278, z = 10.309, p < .001$). Unlike Experiment 1, there was also a significant effect of context timing, such that the context-after group had overall lower /p/ responses ($\hat{\beta} = -.476, z = -2.233, p = .026$). Additionally, there was a marginally significant interaction between context and context timing, such that the context effect was smaller in the context-after group ($\hat{\beta} = -.735, z = -1.771, p = .077$).

3.3. Discussion. In Experiment 2, we successfully replicated the effect of speech adaptation with sentence context as a cue to lexical (and thus phonemic) identity (Jesse 2021, 2024; Aoki & Zellou 2025). However, the perceptual learning effect was relatively weak: while there was

an overall effect of shift condition in the test phase, and no interaction between context timing groups, the simple effects revealed a muddier picture. In the simple effects model, the effect of shift condition was only significant in the context-after condition. Therefore, it is difficult to tell whether the context timing conditions resulted in different degrees of learning.

The relatively small differences between shift conditions appear to be driven by a quirk of our results: participants' /p/-responses were generally much higher than in Experiment 1 and in our norming study. While we estimated a category boundary of 25ms in our norming study, the average participant's category boundary in Experiment 2 was around 15ms, with participants reaching a ceiling of /p/ responses around 30-35ms VOT.[3] When responses are close to ceiling, it becomes more difficult to statistically distinguish differences between shift conditions.

One possible way for us to mitigate this issue would be to only analyze VOT points that are not near the floor or ceiling of responses; however, we would likely lose even more statistical power by eliminating two-three out of our seven VOT steps. Instead, we conduct a third experiment where we will present participants only with VOT steps of 10, 20, 25, and 30ms; this corresponds to intermediate responses in the test phase of Experiment 2 (approximately 25-75% /p/ responses.) We then repeat these intermediate steps more times in order to achieve higher statistical power in this region.

4. Experiment 3.

4.1. Methods.

4.1.1. Participants.

160 subjects were recruited via Prolific and were compensated \$6.00[4] for their participation. All participants were native speakers of American English currently living in the United States who did not participate in Experiments 1-2. As with Experiments 1-2, we removed all participants with below 80% accuracy on filler trials. 13 participants were removed, leaving us with 147 participants whose data were analyzed.

4.1.2. Materials.

Sentence materials were identical to Experiments 1 and 2.

4.1.3. Procedure.

The procedures were identical to Experiment 2 with one minor change: participants were tested at VOTs of 10, 20, 25, and 30ms; each step was repeated four times for each target word, for a total of 48 test trials.

4.1.4. Analyses.

Analyses were identical to Experiment 2.

4.2. Results.

4.2.1. Test phase.

Figure 5 shows the test phase results by shift and context timing condition. As in Experiment 1, we observed a significant effect of VOT on /p/-responses ($\hat{\beta} = 2.192, z = 30.013, p < .001$). Like Experiment 2, we observe evidence for perceptual learning: there was a significant effect of shift condition on /p/-responses, such that right-shifted participants had a lower proportion of /p/-responses, consistent with a right-shifted category boundary ($\hat{\beta} = -.709, z = -2.426, p = .015$). There was not a significant interaction between shift condition and context timing ($p = .302$). The simple-effects model revealed that in the context-before

[3] We postpone further discussion of why there might be differences in categorization behavior to the General Discussion.

[4] This experiment was conducted several months after Experiments 1-2 and was based on an updated compensation rate of \$12.00 per hour.

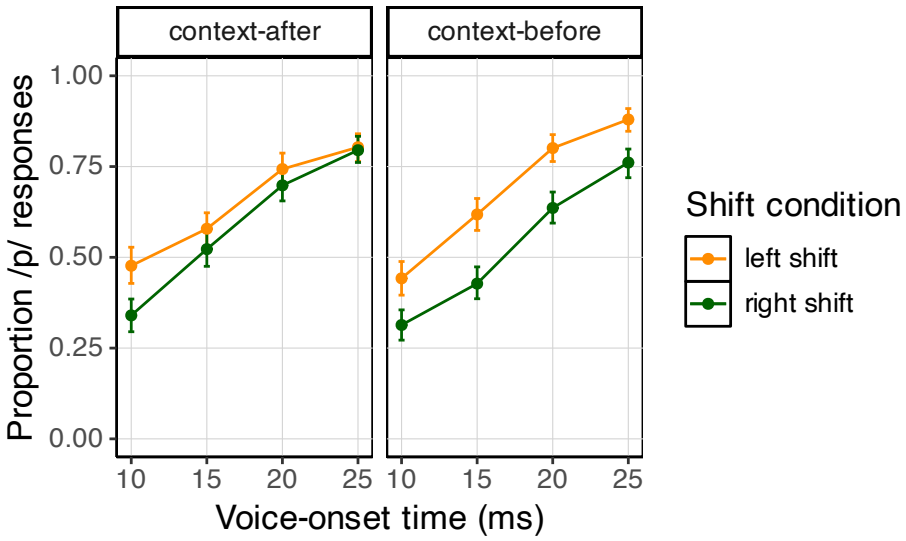


Figure 5. Experiment 3 /p/ responses (y-axis) during the test phase by VOT (x-axis), shift condition (color) and context timing (panel). Error bars are 95% confidence intervals bootstrapped over subject means.

condition, there was a significant effect of shift group ($\hat{\beta} = -1.01, z = -2.459, p = .014$); but there was no significant effect of shift condition in the context-after group ($\hat{\beta} = -.407, z = -.981, p = .327$).

4.2.2. EXPOSURE PHASE. In the exposure phase, there was a main effect of context such that /p/-biased context sentences led to higher /p/ responses ($\hat{\beta} = 6.47, z = 23.48, p < .001$). There was also a main effect of shift condition, such that the right-shifted group had higher /p/ responses ($\hat{\beta} = 2.047, z = 10.119, p < .001$). There was a marginally significant main effect of context timing, such that the context-after group had overall lower /p/ responses ($\hat{\beta} = -0.339, z = -1.705, p = .088$). There were no other significant effects.

4.3. DISCUSSION. Experiment 3 is qualitatively very similar to Experiment 2: we observed overall recalibration effects, no interactions between context timing groups, but we again observed differences in the simple effects. Here, there was only a significant effect of shift condition in the context-*before* group, as opposed to Experiment 2 where we only observed a significant effect in the context-*after* group.

5. General Discussion. In three experiments, we aimed to test whether sentence context occurring before or after a manipulated target word can induce perceptual recalibration. We found no evidence of learning in either sentence context timing group in Experiment 1; however, participants were tested on stimuli where the target contrast occurred in a different phonetic context than in the exposure phase. In Experiments 2 and 3, we fixed this issue and did observe significant learning. While there was no interaction with sentence context timing in either experiment, a simple effects analysis revealed inconsistencies between the two experiments: in Experiment 2, only the context-after group showed a significant learning effect, but in Experiment 3 the reverse held.

The relatively weak learning effects we observe here likely stem from low statistical power.

There are three potential issues at play. The first is our low number of test trials (48-70 between experiments). There is a principled reason to keep the test phase of adaptation studies short: because the test stimuli are uniformly distributed, it can cause participants to un-learn what they learned in exposure. We kept our test phase as short as possible, and in line with previous perceptual recalibration experiments using sentence context, but having fewer trials will always mean lower statistical power. One way to mitigate this problem in future work is to employ a repeated exposure-test paradigm, keeping individual test phases short (see Vroomen et al. 2007). On its own, our short test phase is unlikely to be a critical issue (previous studies successfully find learning effects with test phases of similar length Jesse 2021, 2024; Aoki & Zellou 2025); however, there are two features of our experimental stimuli that likely also affected our power to detect learning effects.

In stimulus design, we generally avoided having other words in our exposure sentences containing /b/ or /p/; however, the exposure sentences on average contain about .5 words with /b/ or /p/ onsets (and this does not account for occurrences at other positions in the words, or for stops at other places of articulation in word-onset position). Given that these other stops were not acoustically manipulated and followed the speaker's natural distribution, this likely counteracted the effects of the shift conditions and led to smaller learning effects. Previous work investigating perceptual recalibration using sentence stimuli were more careful to avoid problematic stimuli in the exposure phase (Jesse 2021, 2024; Aoki & Zellou 2025). In our view, the fundamental problem is that these stimuli were originally developed for spoken word recognition experiments (Westwig et al. under review) where this property is generally non-problematic; however, they may be inappropriate for perceptual recalibration experiments. While it is difficult to avoid this problem entirely (many English words begin with /b/ or /p/), it can be more carefully controlled; researchers can also utilize recent computational frameworks like Xie et al. (2023) to predict how their exposure stimuli will affect perceptual learning.

A final issue related to our stimuli is the overall shift in test phase categorization curves we observed in Experiments 2-3. As we noted, we conducted a norming study on our target words isolated from sentences and found a category boundary of about 25ms VOT. However, in the test phase of Experiments 2-3, when participants hear the exact same stimuli as those in the norming study, the categorization curves shifted quite dramatically to the lower end of the VOT spectrum (with category boundaries as low as 15ms). This overall shift resulted in many test trials being at ceiling, making it hard to detect differences between shift conditions. Furthermore, this unpredictable shift in behavior makes interpreting any learning effects that *are* present more difficult. While we are not certain of the cause of this shift, one possibility is difference in perceived speech rate. It is possible that participants perceived the speech rate of the isolated words as faster compared to the speech rate of the words when embedded in full sentences. Faster speech rates affect the perception of time-based cues like VOT, making shorter VOTs seem perceptually longer (i.e., more /p/-like). This would explain why participants' category boundaries were shifted toward the lower end of the VOT continuum. It would also explain why we observed differences between the norming study and the present experiments—in the norming study, participants never heard the words in the context of full sentences, so they did not have the same point of reference for speech rate. This leaves open why other experiments using full sentences have not observed this issue (Jesse 2021, 2024; Aoki & Zellou 2025), and why our participants in Experiment 1 showed the expected categorization curves. In all of those studies, however, participants are tested on non-word syllables, which may behave differently than our test stimuli which

appear in both exposure and test.

With these issues in mind, it is difficult to come to any firm conclusions about the present experiments. Despite all the problems noted above, however, we did observe consistent learning effects in Experiments 2-3; this adds to the growing literature showing that sentence context can aid in perceptual recalibration (Jesse 2021, 2024; Aoki & Zellou 2025). Furthermore, in one of our experiments, the group exposed to sentence context after the target word showed a significant learning effect. This suggests that at present, we can't rule out that sentential context arriving several syllables beyond a target word can promote perceptual recalibration. If these results hold in more carefully controlled future work, it would suggest that listeners may be capable of maintaining perceptual representations of past speech significantly beyond the word boundary.

References

Andruski, Jean E, Sheila E Blumstein & Martha Burton. 1994. The effect of subphonetic differences on lexical access. *Cognition* 52(3). 163–187. https://doi.org/10.1016/0010-0277(94)90042-6.

Aoki, Nicholas & Georgia Zellou. 2025. When multiple talker exposure is necessary for cross-talker generalization: Insights into the emergence of sociolinguistic perception. *Glossa Psycholinguistics* 4(1). https://doi.org/10.5070/g6011.21217.

Barr, Dale J, Roger Levy, Christoph Scheepers & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bicknell, Klinton, Wednesday Bushong, Michael K Tanenhaus & T Florian Jaeger. 2025. Maintenance of subcategorical information during speech perception: Revisiting misunderstood limitations. *Journal of Memory and Language* 140. 104565. https://doi.org/10.1016/j.jml.2024.104565.

Brown-Schmidt, Sarah & Joseph C Toscano. 2017. Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience* 32(10). 1211–1228. https://doi.org/10.1080/23273798.2017.1325508.

Burchill, Zachary, Linda Liu & T Florian Jaeger. 2018. Maintaining information about speech input during accent adaptation. *PloS One* 13(8). e0199358. https://doi.org/10.1371/journal.pone.0199358.

Caplan, Spencer, Alon Hafri & John C Trueswell. 2021. Now you hear me, later you dont: The immediacy of linguistic computation and the representation of speech. *Psychological Science* 32(3). 410–423. https://doi.org/http://dx.doi.org/10.1177/0956797620968787.

Christiansen, Morten H & Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* 39. https://doi.org/10.1017/s0140525x1500031x.

Connine, Cynthia M. 1987. Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language* 26(5). 527–538. https://doi.org/10.1016/0749-596x(87)90138-0.

Connine, Cynthia M, Dawn G Blasko & Michael Hall. 1991. Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language* 30(2). 234–250. https://doi.org/10.1016/0749-596x(91)90005-5.

Davis, Matthew H & Ingrid S Johnsrude. 2007. Hearing speech sounds: Top-down influences on

the interface between audition and speech perception. *Hearing Research* 229(1-2). 132–147. https://doi.org/10.1016/j.heares.2007.01.014.

Falandays, J Benjamin, Sarah Brown-Schmidt & Joseph C Toscano. 2020. Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language* 112. 104088. https://doi.org/10.1016/j.jml.2020.104088.

Ganong, William F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6(1). 110–125. https://doi.org/10.1037//0096-1523.6.1.110.

Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27(15). 2865–2873. https://doi.org/10.1002/sim.3107.

Jesse, Alexandra. 2021. Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47(1). 184–194. https://doi.org/10.1037/xlm0000805.

Jesse, Alexandra. 2024. Phonetic retuning to idiosyncrasies in word onsets: The interplay of lexical context and prediction. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 50(12). 19181931. https://doi.org/10.1037/xlm0001411.

Kleinschmidt, Dave F. 2019. Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience* 34(1). 43–68. https://doi.org/10.1080/23273798.2018.1500698.

Kleinschmidt, Dave F & T Florian Jaeger. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122(2). 148–203. https://doi.org/10.1037/a0038695.

Lisker, Leigh & Arthur S Abramson. 1967. Some effects of context on voice onset time in English stops. *Language and Speech* 10(1). 1–28. https://doi.org/10.1177/002383096701000101.

Lisker, Leigh & Arthur S Abramson. 1970. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*, vol. 563, 563–567. Academia Prague Prague, Czech Republic.

Luthra, Sahil, James S Magnuson & Emily B Myers. 2021. Boosting lexical support does not enhance lexically guided perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47(4). 685–704. https://doi.org/10.1037/xlm0000945.

McMurray, Bob, Michael K Tanenhaus & Richard N Aslin. 2009. Within-category VOT affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language* 60(1). 65–91. https://doi.org/10.1016/j.jml.2008.07.002.

Norris, Dennis, James M McQueen & Anne Cutler. 2003. Perceptual learning in speech. *Cognitive Psychology* 47(2). 204–238. https://doi.org/10.1016/s0010-0285(03)00006-9.

R Core Team. 2024. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`.

Vroomen, Jean, Sabine van Linden, Béatrice De Gelder & Paul Bertelson. 2007. Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia* 45(3). 572–577.

Westwig, Anna, Yoolim Kim & Wednesday Bushong. under review. Acoustic and semantic cues are optimally combined in spoken word recognition: Evidence from perception in clear and noisy speech .

Winn, Matthew B. 2020. Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *The Journal of the Acoustical Society of America* 147(2). 852–866. https://doi.org/10.1121/10.0000692.

Xie, Xin, T. Florian Jaeger & Chigusa Kurumada. 2023. What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex* 166. 377424. https://doi.org/10.1016/j.cortex.2023.05.003.

Xie, Xin & Chigusa Kurumada. 2025. Phonetic cue distributions guide perceptual adaptation in speech: Evidence from a three-week study with a natural non-native accent. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 47, 1260–1267. https://escholarship.org/uc/item/4p91z4vq.