frontiers | Frontiers in Communication

ORIGINAL RESEARCH
published: 24 May 2022
doi: 10.3389/fcomm.2022.858520

# Immediate Integration of Coarticulatory Cues for /s/-Retraction in American English

*Jacob B. Phillips**

*Department of Linguistics, University of Chicago, Chicago, IL, United States*

Coarticulatory "noise" has long been presumed to benefit the speaker at the expense of the listener. However, recent work has found that listeners make use of that variation in real time to aid speech processing, immediately integrating coarticulatory cues as soon as they become available. Yet sibilants, sounds notable for their high degree of context-dependent variability, have been presumed to be unavailable for immediate integration, requiring that listeners hold all cues in a buffer until all relevant cues are available. The present study examines the cue integration strategies that listeners employ in the perception of prevocalic and pre-consonantal sibilants. In particular, this study examines the perception of /s/-retraction, an ongoing sound change whereby /s/ is realized approaching /ʃ/ as a result of long distance coarticulation from /r/. The study uses eye tracking in the Visual World Paradigm in order to determine precisely when listeners are able to utilize the spectral cues in sibilants in different phonological environments. Results demonstrate that while in most instances listeners wait until more cues are available before considering the correct candidate, fixation accuracy increases significantly throughout the sibilant interval alone. In the pre-consonantal environment, immediate integration strategies were strengthened when the coarticulatory cues of retraction were stronger and when they were more predictable. These findings provide further evidence that context-dependent variation can be helpful to listeners, even on the most variable of sounds.

**Keywords: speech perception, sound change, cue integration, ambiguity, sibilants, coarticulation**

OPEN ACCESS

***Edited by:***
*Georgia Zellou, University of California, Davis, United States*

***Reviewed by:***
*Jonathan Manker, Rice University, United States*
*Navin Viswanathan, The Pennsylvania State University (PSU), United States*

***Correspondence:**
*Jacob B. Phillips*
*jbphillips@uchicago.edu*

***Specialty section:***
*This article was submitted to Frontiers in Communication, a section of the journal Frontiers in Communication*

***Received:*** *20 January 2022*
***Accepted:*** *28 April 2022*
***Published:*** *24 May 2022*

***Citation:***
*Phillips JB (2022) Immediate Integration of Coarticulatory Cues for /s/-Retraction in American English. Front. Commun. 7:858520. doi: 10.3389/fcomm.2022.858520*

## 1. INTRODUCTION

Coarticulation has often been considered to be a process that primarily aids the speaker, as it decreases the articulatory distance between two adjacent gestures and may therefore decrease articulatory effort (Lindblom, 1990). Coarticulation can work in both directions, with preceding sounds affecting following sounds (carry-over coarticulation) and following sounds affecting preceding sounds (anticipatory coarticulation). For some researchers, coarticulation has been viewed as a process that not only aids the speaker, but also actively hinders the listener, as the increased degree of coarticulation between gestures may render the speech signal more ambiguous (Stevens and Keyser, 2010). Under this view, phonetic ambiguity arises because the coarticulated speech deviates substantially from the citation form,

which in turn may diminish potential phonological contrasts between two sounds. Such accounts propose that in listener-directed speech, speakers will minimize coarticulation, thereby increasing articulatory effort and thus consequentially avoiding any potential ambiguity that could inhibit listener comprehension. However, research on elicited clear speech has found that speakers do not reduce anticipatory coarticulation in clear speech compared to normal conditions (Matthies et al., 2001). Other work has demonstrated that coarticulation is increased, rather than decreased, for more confusable words, suggesting that coarticulation itself, rather than the reduction of it, may be a form of hyperarticulation (Scarborough, 2004). This finding holds for different languages (English vs. French), different directions (anticipatory vs. carryover), and different types of coarticulation (vowel-to-vowel coarticulation, which could potentially reduce a phonemic contrast, and vowel nasalization, which would not imperil a phonemic contrast).

In this vein, many approaches to coarticulation propose that it is a process that mutually aids both speaker and listener. That is, while coarticulation may result in diminished phonological contrasts and greater deviation from citation forms, it provides listeners with helpful contextual information from adjacent phones, potentially easing the perception of the sounds in their relevant contexts. Treating coarticulation as a process that creates ambiguity disregards the role of context: What is ambiguous in isolation is not only clear, but beneficial, in context. This approach is built into varying, and often times, conflicting models of speech perception. Gesturalists posit that successful speech perception is accomplished by recovering the articulatory gestures that the speaker produced (Fowler, 1986, 1996, 2006) or intended (Liberman and Mattingly, 1985). In a gesturalist account, listeners make use of coarticulatory variation in order to better recover those gestures (e.g., Viswanathan et al., 2010). In contrast, auditorist approaches posit that listeners rely exclusively on their fine-tuned auditory systems and need not recruit their experiences as speakers (Lotto and Kluender, 1998; Diehl et al., 2004). In an auditorist account, our general auditory systems are sufficiently developedto account for and utilize context-dependent variation based off the acoustic signal alone (Lotto and Kluender, 1998; Holt and Kluender, 2000). Yet while these theories have much they disagree on, both approaches agree that coarticulation is more than something than can be overcome—it provides useful context-dependent information that aids, rather than hinders, speech perception[1]. Similarly, models of speech perception, like TRACE, also incorporate the perceptual benefit of coarticulation in word recognition (Elman and McClelland, 1986).

This perceptual benefit of coarticulation has been demonstrated robustly in the laboratory. Listeners are able to correctly identify the target word more quickly and accurately when more coarticulatory information is present (Martin and Bunnell, 1981; Whalen, 1991; Connine and Darnieder, 2009). Similarly, listeners are more accurate in identifying deleted segments when coarticulatory information is present than when it is missing (Ostreicher and Sharf, 1976). The development of eye-tracking has allowed researchers to examine the perceptual benefit of coarticulation in real-time, asking not only how contextual information improves task accuracy, but also how listeners use the cues of coarticulation to anticipate upcoming sounds. For example, Beddor et al. (2013) examined the perception of anticipatory nasal coarticulation, presenting listeners with two pictures that varied only on the presence or absence of the nasal consonant, e.g., *scent* /sɛnt/ and *set* /sɛt/. Beddor et al. (2013) found that listeners can anticipate the upcoming nasal, looking to an image like *scent* off coarticulation alone even before the nasal consonant is heard. However, the absence of nasality was not equally helpful; that is, oral vowels did not lead to faster or more accurate looks to words like *set*. These findings not only bolster earlier behavioral accounts that coarticulatory information is helpful to the listener, but also show that listeners can use that information as soon as it becomes available. This process by which listeners immediately use available information in lexical identification has been referred to as immediate integration or a "cascade" perception strategy. In addition to nasalization, immediate integration has been demonstrated for a variety of contrasts in which cues become available sequentially, like stop voicing (McMurray et al., 2008).

In contrast, a "buffer" strategy or delayed integration strategy describes the process by which listeners hold the unfolding information in a buffer until all relevant cues are available before beginning lexical identification. Galle et al. (2019) have suggested that, unlike for stops and nasalization, listeners use a buffer strategy for sibilant perception. That is, despite the potential for listeners to use spectral cues to immediately distinguish sibilants like /s/ and /ʃ/, the primary cues in contrasting the two places of articulation, listeners wait for the formant transitions, a secondary cue. Galle et al. (2019) explored a variety of possible explanations for this observation ranging from an auditory account that sibilants make contrasts at higher frequencies than other sounds to the possibility that spectral cues in sibilants are not reliable enough or simply too context-dependent and variable. The latter hypothesis is of particular interest as it contradicts findings of immediate integration for coarticulation like Beddor et al. (2013), which illustrate that context-dependent variation in vowels can be immediately integrated and help anticipate upcoming sounds due to the structured and predictable nature of coarticulation. The present study puts these different accounts in conversation through an examination of cue integration strategies for sibilant coarticulation. In particular, this study examines sibilant coarticulation in preconsonantal environments where coarticulation is predictable, but no formant transitions are available such that listeners could rely on those potential secondary cues.

The focus of the present study is /s/-retraction, a sound change in progress in many varieties of English by which /s/ approaches /ʃ/ in the context of /r/, most notably in /str/ clusters[2] So for a speaker exhibiting /s/-retraction, a word like *street* /strit/ may sound more like *shtreet* /ʃtrit/. This

[1]For a recent review the role of context-dependent perception in gesturalist and auditorist approaches (see Stilp, 2019).

[2]For a detailed discussion of the production, perception, and phonological accounts of /s/-retraction (see Phillips, 2020).

has been observed in various dialects of American English (Shapiro, 1995; Durian, 2007; Baker et al., 2011; Gylfadottir, 2015; Wilbanks, 2017; Smith et al., 2019; Phillips, 2020) as well as varieties of English across the Anglophone world (Lawrence, 2000 for New Zealand; Glain, 2013; Bailey et al., 2022 for the United Kingdom; Stevens and Harrington, 2016 for Australia). Additionally, corpus studies have demonstrated that /s/-retraction is advancing in apparent time in the United States (Gylfadottir, 2015; Wilbanks, 2017). At its core, /s/-retraction can be viewed as a coarticulatory process by which /s/ is produced with greater tongue body retraction and lip protrusion so as to minimize articulatory distance between /s/ and /r/ (Baker et al., 2011; Smith et al., 2019). These small articulatory changes can have outsized acoustic effects, resulting in a sibilant more characteristic of an /ʃ/ than /s/ (Baker et al., 2011). However, despite resulting in a sibilant that may surface between /s/ and /ʃ/, /s/-retraction need not necessarily create confusion due to the phonotactic restrictions of English: While /s/ and /ʃ/ are contrastive prevocalically, only /ʃ/ precedes /r/ and only /s/ precedes all other consonants. Thus, English phonotactic restrictions on preconsonantal sibilants create an environment in which extreme coarticulation is unfettered by potential lexical confusability.

In order to address these notions of ambiguity and confusability, the perception of /s/-retraction, not just its production, needs to be examined, and while a growing body of work has examined the production of the sound change, scant work has examined listeners' perception of it. In one perception study, Kraljic et al. (2008) found that exposure to sibilants ambiguous between /s/ and /ʃ in /str/ clusters, like *industry* /ɪndəʃtri/, where retraction is expected, does not alter an individual's /s/-/ʃ/ categorization as strongly as ambiguous sibilants in unpredictable prevocalic environments, like *dinosaur* /daɪməʃɔr/. In another, Phillips and Resnick (2019) examined the perception of onset sibilants in nonce words, like *strimble* or *shtrimble*, where listeners may be less constrained by lexical/phonotactic restrictions. Phillips and Resnick (2019) found that individuals were less categorical, and less likely to perceive an /ʃ/ onset in /str/ clusters, where /s/-retraction is more expected, than in /spr/ and /skr/ clusters. Both studies demonstrate that listeners have detailed context-dependent knowledge about /s/-retraction based off their experiences. The present study asks how listeners use that information in real time. That is, can listeners use their knowledge of context-dependent spectral variation in sibilants in order to more quickly and accurately identify the target word? And crucially, by looking at perception in real time, we can examine how listeners deal with a case of ephemeral ambiguity: The ambiguity between the sibilants in these environments exists only for a short amount of time until disambiguating information, like the ultimate presence or absence of /r/, follows. Additionally, through an examination of a sound change in progress, rather than a potentially more stable coarticulatory pattern like vowel nasalization, the present study builds on previous work on cue integration to ask whether listeners are consistent and uniform in their use of a changing cue.

# 2. METHODS

## 2.1. Participants

A total of 52 participants were recruited from the University of Chicago undergraduate subject pool and received course credit or payment. All participants were between 18 and 22 years of age. Thirty-seven participants identified as female, 15 as male, and none as non-binary or transgender. Just over half of the participants (29) identified as straight/heterosexual. Similarly, 29 participants identified as white. Participants were geographically distributed across the United States, with more participants reporting growing up in suburban areas (34) compared to urban (15) or rural (3) environments. All participants were self-reported native speakers of North American English with no history of hearing loss, language and communication disorders, or any other medical conditions commonly associated with cognitive impairment. An additional nine individuals participated in this study but were excluded from analysis due to non-native status, language or neurological disorders, and/or non-attentive responses.

## 2.2. Stimuli

The target stimuli were designed to manipulate the degree of retraction in sibilant clusters to examine whether the anticipatory cues of /r/ presence can influence lexical processing. The stimuli thus included the relevant /sCr/ and /sC/ clusters as well as simplex prevocalic /s/ and /ʃ/. There were three sets of near minimal pair quadruplets, one for each place of articulation of the intervening stop: *sit-spit-spritz-shit* (bilabial), *sing-sting-string-shingle* (alveolar), and *sip-skip-script-ship* (velar). Stop initial quadruplets also varying in place of articulation and presence of /r/ were included as fillers: *pick-prick-brick-big* (bilabial), *tip-trip-drip-dip* (alveolar), and *kit-crypt-grip-gift* (velar).

The original auditory stimuli were produced by a college-aged male from Illinois. The speaker recorded five repetitions of each target word in the carrier phrase "Now select X." All stimuli materials were recorded at 48,000 Hz with a Shure SM10A head-mounted microphone in a sound-attenuated booth.

To provide control and consistency over the degree of retraction in the onset sibilants, all stimuli were cross-spliced. The onset sibilants from the target words were deleted and replaced with a sibilant digitally mixed from prevocalic /s/ (*sip*) and /ʃ/ (*ship*) at different scaling ratios, using a Praat script originally created by Darwin (2005). For each /sCr/ cluster, three degrees of retraction were used to test the hypothesis that listeners attend to coarticulation on the sibilant to anticipate the presence of absence of an upcoming /r/: minimal, moderate, and extreme retraction. The retraction conditions were designed in consultation with previous examinations of /s/-retraction (e.g., Baker et al., 2011), with the talker's natural production of /s/ in these environments, and with the researcher's perception. In all /sCr/ clusters, the minimal retraction condition was designed to exhibit less retraction than the speaker produces naturally and to be perceived clearly as an /s/; the stimuli was digitally mixed with 30% /ʃ/ and 70% [s] for /str/ clusters and 10% /ʃ/ and 90% [s] for

**TABLE 1 |** Scaling factors used in stimuli creation.

| | Minimal retraction | | Moderate retraction | | Extreme retraction | |
|---|---|---|---|---|---|---|
| | /s/ | /ʃ/ | /s/ | /ʃ/ | /s/ | /ʃ/ |
| /spr/ | 0.90 | 0.10 | 0.60 | 0.40 | 0.30 | 0.70 |
| /str/ | 0.70 | 0.30 | 0.40 | 0.60 | 0.10 | 0.90 |
| /skr/ | 0.90 | 0.10 | 0.60 | 0.40 | 0.30 | 0.70 |

| Across conditions | | |
|---|---|---|
| | /s/ | /ʃ/ |
| /s/ | 1.00 | 0.00 |
| /sp/ | 0.90 | 0.10 |
| /st/ | 0.90 | 0.10 |
| /sk/ | 0.90 | 0.10 |
| /ʃ/ | 0.00 | 1.00 |

**TABLE 2 |** Pairing of visual images organized by place of articulation and onset environment.

| | s–ʃ | s– sC | sC–sCr | sCr–ʃ |
|---|---|---|---|---|
| /p/ | Sit–shit | Sit–spit | Spit–spritz | Spritz–shit |
| /t/ | Sing–shingle | Sing–sting | Sting–string | String–shingle |
| /k/ | Sip–ship | Sip–skip | Skip–script | Script–ship |
| | **T–D** | **T–Tr** | **Tr–Dr** | **Dr–D** |
| /p/ | Pick–big | Pick–prick | Prick–brick | Brick–big |
| /t/ | Tip–dip | Tip–trip | Trip–drip | Drip–dip |
| /k/ | Kit–gift | Kit–crypt | Crypt–grip | Grip–gift |

/spr/ and /skr/ clusters. The moderate retraction condition was designed to exhibit increased degrees of retraction to the model talker's natural production and to be perceived approaching the /s/-/ʃ/ boundary; the stimuli digitally mixed with 60% /ʃ/ and 40% [s] for /str/ clusters and 40% /ʃ/ and 60% [s] for /spr/ and /skr/ clusters. Finally, the extreme retraction condition was designed to contain twice again as much retraction as the speaker produced naturally and be perceived clearly as an /ʃ/; the onsets digitally mixed with 90% /ʃ/ and 10% [s] for /str/ cluster and 70% /ʃ/ and 30% [s] for /spr/ and /skr/ clusters. The /sC/ onsets did not differ between conditions, and digitally mixed with 10% /ʃ/and 90% [s] in the minimal, moderate, and extreme retraction conditions, consistent with the talker's natural production. So that all stimuli underwent similar manipulations, the onset sibilants in prevocalic environments were also cross-spliced; however, they were not digitally mixed since no retraction would be expected prevocalically. The scaling factors used for the creation of each onset environment can be seen in **Table 1**. Furthermore, to reduce the effects of stimuli manipulation, the stop-initial fillers were cross-spliced with onsets containing manipulated degrees of aspiration.

For each target word, four free and publicly available clipart images were selected, resized, and gray-scaled. Four naïve volunteers selected the image that best corresponded the intended word. In order to control for differences of style, darkness, or image resolution, all images were redrawn by hand, making adjustments to remove any text or distracting features. The hand-drawn images were then scanned, gray-scaled, and resized to 550 × 550 pixels.

## 2.3. Procedure

After informed consent, participants were first familiarized with the images and their associated lexical items. This was more straightforward for nouns and high frequency words than for adjectives, verbs, and low frequency items. Participants were first introduced to the images and their accompanying orthographic labels in a randomized order. Participants were asked to read the label aloud and explain to the researcher how the label relates to the image. To explain the task, the researcher provided two examples verbally: for a picture of a dog with the label "dog," the researcher would simply say "this is a dog," but for a picture of a cheetah with a label "fast," the researcher would say "cheetahs are fast." Following this connection-making task, participants were then shown images in a randomized order without the accompanying orthographic labels and asked to reproduce the corresponding label. All participants exhibited 100% accuracy in the label reproduction task, demonstrating that they had successfully associated the lexical items with the images. No subsequent effect of grammatical category or lexical frequency was observed.

For the identification task, participants were randomly assigned to one of three retraction conditions: minimal, moderate, or extreme retraction. Participants were seated in front of a Tobii T-60 eye-tracker, with a sampling rate of 60 Hz that was recalibrated for each participant. Two images, rather than the typical four, were presented in each trial in a modified Visual World Paradigm (Allopenna et al., 1998). This modification, in which only a single target and competitor image are presented without distractors, was also utilized by Beddor et al. (2013) for an examination of cue integration strategies for anticipatory nasalization. It should be noted that this modification may increase the sensitivity and likelihood that participants will exhibit looks to the target image sooner, centering the question of *can* listeners immediately use the spectral cues of sibilants rather than do they necessarily use them in normal conversations. The images were paired according to contrasts in **Table 2**, with the critical pair for the present study being /s/ vs. /ʃ/, e.g., *sing* vs. *shingle*, and /sC/ vs. /sCr/, e.g., *sting* vs. *string*. Thus, in each trial, participants were only considering one potential sibilant contrast, either a phonemic contrast between /s/ and /ʃ/ or context-dependent variation within a category. Participants were first asked to scan the screen and, after identifying the images, focus on a fixation cross in the center of the screen, equidistant between both images. Once a fixation on the cross was detected, a red box was displayed surrounding the cross. Participants were able to click on the box to play the auditory stimuli "Now select [word]," e.g., "Now select *sting*." Participants were directed to click on the corresponding image as quickly as they could, which signaled the end of the trial and automatically advanced to the next item. Each trial lasted roughly 5 s. Left and right eye movement was recorded throughout the experiment. A sample trial slide is provided in

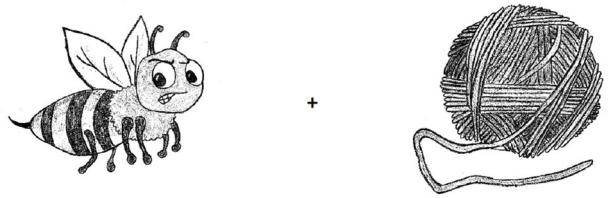

**FIGURE 1 |** A sample trial illustrating the presentation of the images, *sting* **(left)** and *string* **(right)**, and fixation cross.

**Figure 1** to illustrate how the visual stimuli and response options were presented.

## 2.4. Measurements

Both accuracy and gaze measurements were collected. Trial accuracy was defined by clicking on the correct image corresponding to the auditory stimuli. Although a trial may be ambiguous during the onset sibilant portion, the ultimate presence or absence of /r/ would disambiguate the stimulus. Thus, all participants exhibited >95% accuracy in image selection.

Participants' eye gaze was monitored from the initial display of the target and competitor images, through the cross fixation, until 2,000 ms following the onset on the target word or until they clicked on an image, whichever came first. Although eye gaze was tracked for both left and right eyes, analysis was conducted on the right eye exclusively. Unlike trial accuracy, which identifies whether the participant selected the correct image corresponding to the auditory stimuli, gaze measurements identify precisely when the target or competitor lexical item were considered, before the ultimate decision to click on the correct image was made. This not only provides a much more fine-grained temporal resolution than reaction time for mouse clicks, but also allows for an examination of alternative phonological candidates for the ultimately unambiguous stimuli.

The online measurement selected for analysis for the present experiment was the proportion of correct fixations over time, which is determined by examining the accuracy of each individual fixation. A fixation was determined to be a correct fixation if the right eye gaze fell within the 550 × 550 pixel region containing the image corresponding to the auditory stimuli. Fixations were binned into 20 ms windows. A proportion of 0 for a given bin means that there were no trials in the relevant condition during which eye gaze was detected within the 550 × 550 pixel region containing the target image. This means that all participants' gaze was directed at the fixation cross, the competitor image, or anywhere else on the screen other than the target image. Thus, it is not the proportion of target versus competitor fixations, but rather the proportion of target versus non-target fixations. Similarly, a proportion of 1 means that in all trials a target fixation was detected within the specified 20 ms window.

## 2.5. Predictions

The specific hypotheses for participants' eye gaze are as follows:

Hypothesis 1 states that listeners will make immediate use of spectral cues to distinguish /s/ and /ʃ/ in prevocalic environments. This hypothesis is formulated in direct response to the buffer strategy observed for prevocalic sibilants by Galle et al. (2019). Under this hypothesis, correct fixations on /s/ or /ʃ/ will emerge during the onset sibilant, when only spectral information can distinguish the two places of articulation. This hypothesis is tested by TIMEWINDOW in /s/-/ʃ/ pairs. If listeners exhibit increased proportion of correct fixations over the sibilant interval, it suggests that they are using a cascade strategy for integrating the spectral cues of the onset sibilants, contra (Galle et al., 2019). If listeners wait until the onset of the vowel to increase their proportion of correct fixations, this suggests that a buffer strategy is used for sibilants. If such a buffer strategy is observed, Hypotheses 2–4 ask if this is true for pre-consonantal sibilants as well as prevocalic sibilants.

Hypothesis 2 states that listeners will make use of the coarticulatory cues in predicting the phonological context of the sibilant and do so as soon as those cues are available. Under this hypothesis, correct fixations on /sC/ or /sCr/ will emerge during the onset sibilant, before the ultimate absence or presence of /r/ disambiguates the stimuli. If such a pattern is observed, this demonstrates that like with vowel-nasal coarticulation observed by Beddor et al. (2013), long distance rhotic-sibilant coarticulation is immediately available and beneficial to the listener. Like in the prevocalic model, this hypothesis is again tested by TIMEWINDOW, but in examination of /sC/-/sCr/ pairs. Additionally, this hypothesis is tested by RETRACTIONCONDITION (minimal, moderate, or extreme) and its interaction with TIMEWINDOW, examining if stronger cues of retraction, and thus stronger cues of coarticulation, increase the proportion of correct fixations over the course of the sibilant. If Hypothesis 2 is confirmed, then the following hypotheses stand to be tested:

Hypothesis 3 states that a retracted /s/ is a better indicator of rhotic presence than a non-retracted /s/ is for rhotic absence. That is, does a more retracted, i.e., more /ʃ/-like, onset predict an /sCr/ cluster better than a less retracted, i.e., more /s/-like, onset predicts an /sC/ cluster. A confirmation of this hypothesis would demonstrate that the cues of /s/-retraction are more useful in speech processing than the absence of such cues, much like the findings of Beddor et al. (2013) that a nasal vowel is a better cue of an upcoming nasal stop than an oral vowel is of an upcoming oral stop. This is tested by CLUSTER in examination of /sC/-/sCr/ pairs and its interaction with TIMEWINDOW, where more correct fixations are predicted for /sCr/ clusters than /sC/ clusters over the course of the sibilant.

Hypothesis 4 states that the cues of /s/-retraction are a better indicator of rhotic presence in /str/ clusters compared to /spr/ and /skr/ clusters. A confirmation of this hypothesis would demonstrate that listeners have detailed phonological knowledge about /s/-retraction as a sound change in progress, with greater degrees of retraction observed in /str/ clusters (Baker et al., 2011), and adjust their expectations accordingly. This hypothesis

is tested by PLACE of articulation (alveolar, bilabial, and velar) in examination of /sC/-/sCr/ pairs and its interaction with TIMEWINDOW and CLUSTER, where more correct fixations are predicted for alveolar clusters than bilabial and velar clusters, particularly in /str/ clusters, over the course of the sibilant. This hypothesis thus requires that listeners not only use phonological knowledge about the upcoming rhotic, but also about the upcoming stop before that stop is perceived.

## 3. RESULTS

The results of this experiment are presented in two sections. First, in Section 3.1, the results from the /s/-/ʃ/ pairs are presented, asking if listeners attend to the spectral cues of the onset sibilants immediately or whether they hold them in a buffer until vocalic information is available. This section tests Hypothesis 1. Secondly, in Section 3.2, the results from the /sC/-/sCr/ pairs are presented, which tests Hypotheses 2–4. These pairs ask whether listeners can use the coarticulatory cues of /s/-retraction immediately to anticipate the presence of an upcoming /r/.

### 3.1. Prevocalic Results

The prevocalic analysis asks if listeners can use spectral cues present over the course of the sibilant in order to correctly identify a prevocalic sibilant /s/ and /ʃ/, distinguishing words like *sip* /sɪp/ vs. *ship* /ʃɪp/. To test this, generalized linear mixed-effects models with a logit link function were fit to the accuracy of a given fixation (1,0) using the `glmer()` function in the `lme4` package (Bates et al., 2015) in `R` (R Core Team, 2015). As it takes ∼200 ms to plan and execute an eye movement and as the sibilant was 180 ms in duration, the model examined eye movements during the 180 ms window that began 200 ms following the onset of the stimulus sibilant. The prevocalic model includes trial ORDER (1–384, scaled), TIMEWINDOW of the sibilant (1–180, binned into 20 ms windows and scaled), and ONSET (/s/ and /ʃ/; treatment-coded with /s/ as base) as fixed effects. RETRACTIONCONDITION (minimal, moderate, and extreme) was not included as the prevocalic onsets were not manipulated between conditions. Self-reported responses for demographic categories like GENDER, SEXUALITY, AGE, and REGION did not reach a significance threshold of 0.05 and were pruned from the final models. Preliminary models for the different onset pairings included all two- and three-way interactions between the fixed effects predictors. All interactions that did not reach a significance threshold of 0.05 were pruned from the final models. Additionally, the preliminary models included maximally specified random effects structures, with by-subject random slopes and intercepts, which were progressively simplified until convergence was achieved. The results of the prevocalic logistic regression are presented in **Table 3**. The inclusion of by-subject random intercepts and by-subject random slopes for trial ORDER and ONSET suggests significant individual variability with respect to these predictors. By-item random slopes and intercepts are not included as there is only one item per onset cluster, given the training and time constraints of the current design.

**TABLE 3 |** Model predictions for all main effects and interactions in fixation accuracy for /s/ vs. /ʃ/ onsets, $N$ = 26,750.

| | Est. | SE | z | p |
|---|---|---|---|---|
| *Intercept* | −0.47 | 0.17 | −2.79 | **0.005**** |
| Order | 0.02 | 0.06 | 0.31 | 0.758 |
| TimeWindow | 0.39 | 0.01 | 27.36 | **<0.001***** |
| Onset-SH | −0.17 | 0.09 | −1.75 | 0.081 |

*A positive value indicates a greater prediction of fixations on the target word. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. All p values less than 0.05 are in bold.*

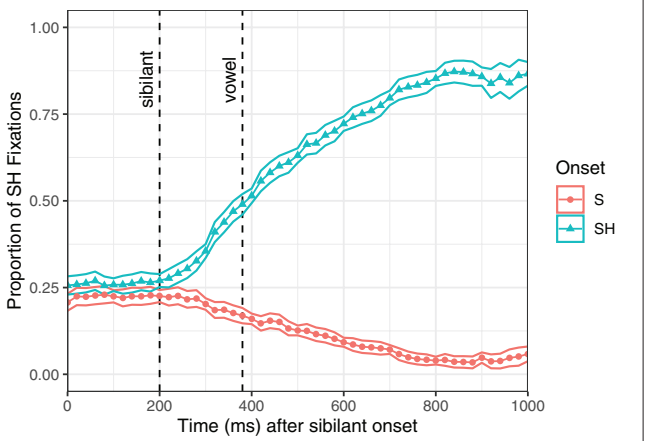


**FIGURE 2 |** Fixation proportion for /ʃ/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 20 ms windows) and onset category (color: /s/ = red circles, /ʃ/ = teal triangles). The vertical lines represent a 200 ms delay from the onset of the sibilant and vowel. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

The negative intercept in the model ($z = -2.79, p = 0.005$) suggests that all else being equal, listeners are more likely to be looking anywhere other than the target image during the sibilant. However, the main effect of TIMEWINDOW ($z = 27.36, p < 0.001$) demonstrates that the proportion of correct fixations increases robustly over the course of the sibilant. The effect of TIMEWINDOW is visualized in **Figure 2**. Although the analysis is conducted on the proportion of correct fixations, I have chosen to visually present the proportion of /ʃ/ fixations. The primary choice in doing so is to allow the fixations for /s/ and /ʃ/ to visually diverge at the time at which the listener's eye gaze between the trials diverges. Unlike in the pre-consonantal stimuli, the prevocalic stimuli are cross-spliced but naturally produced, such that they potentially may be immediately disambiguated. Recall that while immediate disambiguation of sibilants has been demonstrated for /s/ and /ʃ/ in a gating task (Galle et al., 2019), immediate disambiguation has not been demonstrated in speech processing using eye tracking.

**Figure 2** illustrates how the proportion of /ʃ/ fixations changes over the course of a trial. A trial with an /s/ onset is presented in red circles and a trial with an /ʃ/ onset is presented in teal triangles. For both /s/ and /ʃ/ onsets, participants begin with around one quarter of the fixations on the /ʃ/ image, which is

supported by the intercept of the model. While the other fixations are not explicitly indicated in **Figure 2**, they may be to the cross equidistant between the images, where a participant's fixation is required to initiate the trial, or to the competing /s/ image. Since it takes ∼200 ms to plan and execute an eye movement, any fixations planned during the sibilant would be observed ∼200 ms later. Vertical lines are provided in **Figure 2** to indicate what sound was heard when a given eye movement was planned. Thus, if a look to the /ʃ/ image is planned during the sibilant it would be observed between the dashed lines. A look once the vowel has been heard and formant transitions, a secondary cue, are available would be observed following the second dashed line.

Preliminary inspection of **Figure 2** may first highlight that the most dramatic differences between the /s/ and /ʃ/ onsets is not observed until well after the vowel onset is heard. This suggests that in many trials, listeners wait until formant transitions are available to correctly identify the target word, keeping with Galle et al. (2019). However, I am primarily concerned with the fixation proportions *during* the sibilant interval, to ask specifically if listeners *can* use the spectral cues of sibilants even if they don't always do so. At the most basic level, this asks if accuracy of fixations increases over the course of the sibilant, which would be indicated by diverging predictions and steep slopes for /s/ and /ʃ/ onsets between the dashed lines. In **Figure 2**, a dramatic rise in proportion of /ʃ/ fixations is observed between the dashed lines for /ʃ/ onsets paralleled with a notable, but less dramatic, fall for /s/ onsets. Furthermore, the confidence intervals for /s/ and /ʃ/ diverge sharply and almost immediately during the sibilant interval. These visual findings are supported by the model with a significant main effect of TIMEWINDOW ($z = 27.36, p < 0.001$), which suggests that the proportion of correct fixations increases over the course of the sibilant. There is no significant effect of ONSET, either as a main effect or in interaction with any other effects, which suggests that listeners are equally accurate in their perception of /s/ and /ʃ/. However, as the inclusion of by-subject random slopes for ONSET improved model likelihood, there may be significant individual variation in the perception of the different sibilants.

## 3.2. Pre-consonantal Results

As the prevocalic analysis demonstrates that listeners are able to immediately use spectral cues to disambiguate two separate sibilants, the pre-consonantal analysis asks if listeners can use those same cues in order to predict the context of the sibilant. In these stimuli, the contrast is not between two phonemes but rather two phonological environments. The pre-consonantal model is fit on the same 180 ms window but for the /sC/–/sCr/ onsets and includes trial ORDER (1–384, scaled), TIMEWINDOW of the sibilant (1–180, binned into 20 ms windows and scaled), CLUSTER (/sC/ and /sCr/; treatment-coded with /sC/ as base), PLACE of articulation (alveolar, velar, and bilabial; Helmert-coded to first compare alveolar to the combined mean of velar and bilabial and then compare velar to bilabial), and RETRACTIONCONDITION (minimal, moderate, and extreme; treatment-coded with minimal as base) as fixed effects. Like with the prevocalic model, all non-significant interactions and predictors were pruned from the final model and random effects structure was progressively simplified until convergence was achieved. Results of the pre-consonantal model are presented in **Table 4**. The inclusion of by-subject random intercepts and by-subject random slopes for TRIALID, PLACE, and CLUSTER suggests significant individual variability with respect to these predictors.

Like in the prevocalic model, the significant negative intercept ($z = -2.37, p = 0.018$) suggests that participants are more likely to look away from the target image than toward it. And like in the prevocalic model, the main effect of TIMEWINDOW suggests that participants are more likely to look to the correct image over the course of the sibilant ($z = 2.48, p = 0.013$). This effect is noticeably smaller and less robust than in the prevocalic environment. While in the prevocalic environment the spectral cues are the primary cues in making the contrast between the two target items, in the pre-consonantal environment, the spectral cues are secondary coarticulatory cues present while the stimuli remain ambiguous until the ultimate presence or absence of /r/ disambiguates the candidates 77 ms after the end of the sibilant.

Fixations for the different retraction conditions, pooled across places of articulation, is illustrated in **Figure 3**. Although the model is fit on the accuracy of fixations, for the ease of visualization, I present the proportion of /sCr/ fixations. Again, vertical lines are provided as guideposts to what sound was heard when the eye movement was planned, including the following stop. In **Figure 3**, /sCr/ fixations rise noticeably in the moderate and extreme RETRACTIONCONDITION over the course of the sibilant, which is indicated by the positive slopes of the teal lines between the dashed vertical lines. Additionally, the proportions of /sCr/ fixations diverge for the moderate and extreme RETRACTIONCONDITION slightly at the end of sibilant period in both conditions, although the most noticeable divergence occurs after the sibilant ends during the stop period. These observations are supported by the interaction of TIMEWINDOW with RETRACTIONCONDITION in the regression, with more correct fixations predicted over the course of the sibilant in moderate and extreme retraction conditions (moderate: $z = 5.07, p < 0.001$; extreme $z = 3.43, p < 0.001$). These findings suggest that individuals are able to use the available coarticulatory cues of /s/-retraction in order to improve correct fixations, well before the onset of the disambiguating /r/. This interaction effect with RETRACTIONCONDITION also explains the relatively smaller main effect of TIMEWINDOW compared to the prevocalic model: While in the prevocalic /s/ and /ʃ/, helpful spectral cues are equally present in all stimuli, in the pre-consonantal stimuli, only few coarticulatory cues are available in the minimal retraction condition.

**Figure 4** breaks down the findings by place of articulation. Visual inspection of the figure indicates a steeper teal line for /sCr/ clusters and divergence of the red /sC/ and teal /sCr/ confidence intervals in the alveolar onsets compared to the bilabial and velar onsets. This is supported by the model with the significant interaction of TIMEWINDOW, CLUSTER (SCR), and PLACE of articulation ($z = 2.82, p = 0.005$). Recall that place of articulation is Helmert-coded so the comparison made here is between alveolar onsets and the

**TABLE 4 |** Model predictions for all main effects and interactions in fixation accuracy for /sCr/ vs. /sC/ onsets, *N* = 27,067.

| | Est. | SE | z | *p* |
|---|---|---|---|---|
| *Intercept* | −0.68 | 0.29 | −2.37 | **0.018*** |
| Order | 0.01 | 0.07 | 0.14 | 0.890 |
| TimeWindow | 0.07 | 0.03 | 2.48 | **0.013*** |
| Condition (Moderate) | −0.22 | 0.39 | −0.57 | 0.569 |
| Condition (Extreme) | −0.71 | 0.36 | −1.98 | 0.053 |
| Cluster (SCR) | −0.38 | 0.17 | −2.19 | **0.028*** |
| Place (1) | −0.12 | 0.19 | −0.64 | 0.522 |
| Place (2) | −0.04 | 0.23 | −0.19 | 0.851 |
| TimeWindow × Condition (Moderate) | 0.19 | 0.04 | 5.07 | **<0.001***** |
| TimeWindow × Condition (Extreme) | 0.11 | 0.03 | 3.43 | **<0.001***** |
| TimeWindow × Cluster (SCR) | −0.03 | 0.03 | −1.11 | 0.265 |
| TimeWindow × Place (1) | −0.03 | 0.04 | −0.81 | 0.418 |
| TimeWindow × Place (2) | 0.05 | 0.05 | 1.08 | 0.278 |
| Cluster (SCR) × Place (1) | 0.14 | 0.11 | 1.24 | 0.213 |
| Cluster (SCR) × Place (2) | −0.25 | 0.13 | −1.89 | 0.060 |
| Cluster (SCR) × Condition (Moderate) | 0.41 | 0.24 | 1.73 | 0.085 |
| Cluster (SCR) × Condition (Extreme) | 0.52 | 0.22 | 2.38 | **0.017*** |
| Place (1) × Condition (Moderate) | 0.25 | 0.27 | 0.91 | 0.361 |
| Place (1) × Condition (Extreme) | 0.15 | 0.25 | 0.62 | 0.532 |
| Place (2) × Condition (Moderate) | −0.07 | 0.33 | −0.22 | 0.823 |
| Place (2) × Condition (Extreme) | 0.26 | 0.30 | 0.86 | 0.392 |
| TimeWindow × Cluster (SCR) × Place (1) | 0.17 | 0.06 | 2.82 | **0.005** |
| TimeWindow × Cluster (SCR) × Place (2) | −0.04 | 0.07 | −0.065 | 0.516 |
| Cluster (SCR) × Place (1) × Condition (Moderate) | −0.64 | 0.16 | −4.00 | **<0.001***** |
| Cluster (SCR) × Place (1) × Condition (Extreme) | 0.14 | 0.15 | 0.96 | 0.337 |
| Cluster (SCR) × Place (2) × Condition (Moderate) | −0.03 | 0.18 | −0.15 | 0.879 |
| Cluster (SCR) × Place (2) × Condition (Extreme) | 0.13 | 0.17 | 0.75 | 0.453 |

*Place is Helmert-coded: Place (1) indicates alveolar compared to the mean of velar and bilabial; Place (2) indicates velar compared to bilabial. A positive value indicates a greater prediction of fixations on the target word.*

*$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. All p values less than 0.05 are in bold.*

combined mean of velar and bilabial onsets. This suggests that listeners improve their consideration of the correct candidate most in /str/ clusters, precisely where /s/-retraction is both most expected and those cues are most available. No four-way interactions between TIMEWINDOW, CLUSTER, PLACE, and RETRACTIONCONDITION emerged as significant such that individuals were influenced most by greater levels of retraction in alveolar clusters. Rather, the results indicate that high degrees of retraction regardless of place of articulation are helpful to the listener and the spectral cues in /str/ clusters, which by nature of the stimuli always contain more cues of retraction than their bilabial and velar counterparts, aids the listener.

Additionally, the model suggests that other effects and interactions that do not have to do with the timing of sibilant can also influence the listener. Specifically, a main effect of CLUSTER emerged (SCR: $z = -2.19, p = 0.028$), such that individuals are less accurate in their consideration of /sCr/ clusters than /sC/ clusters across the board. This effect is counteracted in the extreme retraction condition by the interaction of CLUSTER (SCR) and RETRACTIONCONDITION (moderate: $z = 1.73, p = 0.085$; extreme: $z = 2.38, p = 0.017$), which suggests that individuals are more accurate in their consideration of /sCr/ candidates when the highest degrees of retraction are available. Finally, a three-way interaction of interaction of CLUSTER (SCR), PLACE of articulation (alveolar compared to the mean of velar and bilabial), and RETRACTIONCONDITION emerged as significant (moderate: $z = -4.00, p < 0.001$; extreme: $z = 0.96, p < 0.337$), such that the beneficial effects of the moderate retraction condition and the alveolar place of articulation are tempered in conjunction with one another.

The models and figures thus far pool data across 52 participants which can potentially obfuscate individual differences in processing styles. That is, we might ask do some participants use a buffer strategy while other participants use those cues more immediately indicative of a cascade strategy? In **Figure 5**, nine individual participants' fixation proportions are visualized, with three participants from each retraction condition. Fixations are pooled across places of articulation and confidence intervals are excluded due to the paucity of observations from a single individual. Participants are categorized into one of three patterns: delayed, buffer, and cascade integration. For participants who exhibit delayed looks,

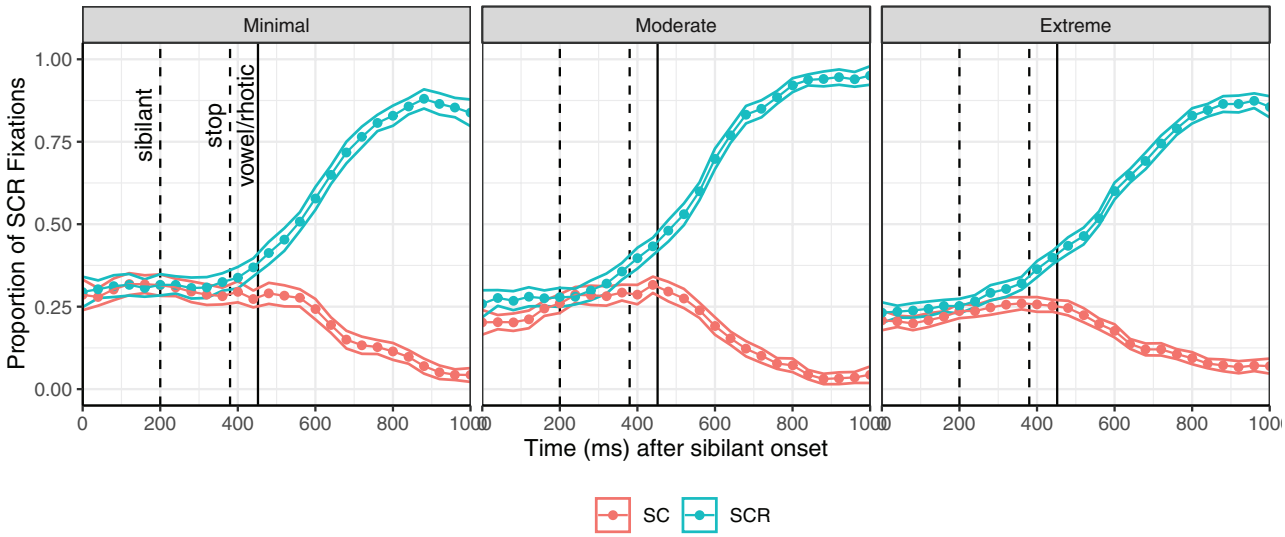


**FIGURE 3 |** Fixation proportion for /sCr/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 40 ms windows), cluster type (color: /sC/ = red circles, /sCr/ = teal triangles), and retraction condition (columns). The vertical lines represent a 200 ms delay from the onset of the sibilant, stop, and vowel/rhotic. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

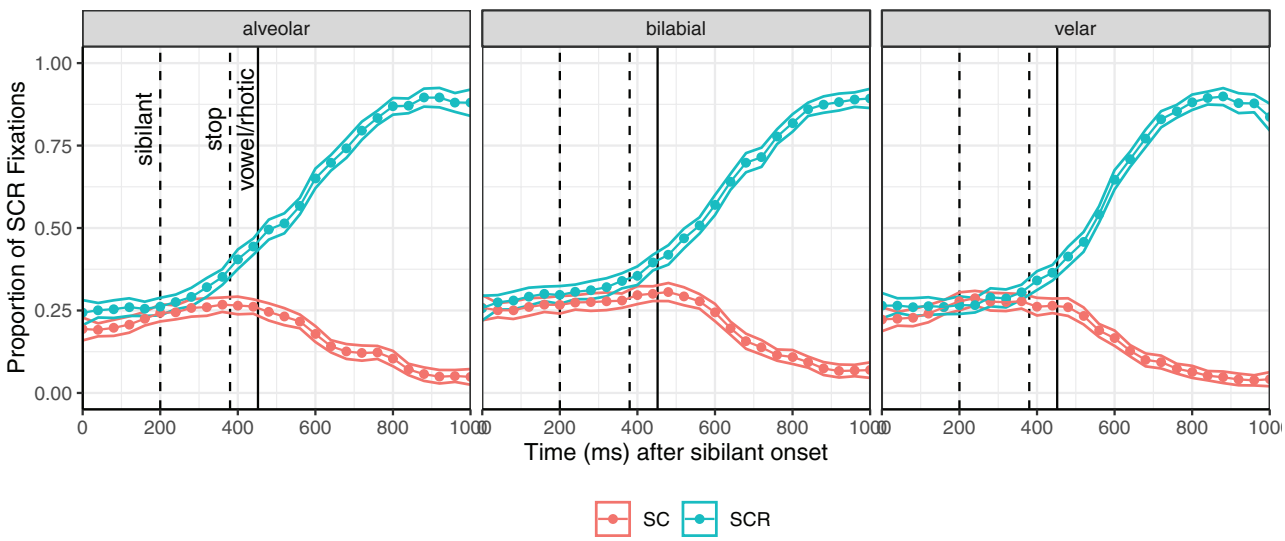


**FIGURE 4 |** Fixation proportion for /sCr/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 40 ms windows), cluster type (color: /sC/ = red circles, /sCr/ = teal triangles), and place of articulation (columns). The vertical lines represent a 200 ms delay from the onset of the sibilant, stop, and vowel/rhotic. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

they begin with near 0% fixations on /sCr/ images, suggesting that they are often maintaining their gaze on the fixation cross, either because they are slower at directing their eye gaze or out of an effort to be a conscientious participant. Participants who exhibit delayed looks thus almost never exhibit clear indications of immediate integration such that their proportion of correct fixations increases during the sibilant interval. A second category is individuals who are looking to either the target or competitor image when the sibilant begins, but their consideration of /sCr/ and /sC/ images do not diverge until the stop or rhotic/vowel portion of the stimuli. These participants appear to exhibit a buffer strategy and wait to integrate the cues of retraction until additional information is available. Finally, the third pattern of participants is individuals who show evidence for increased consideration of the correct candidate during the sibilant portion alone, integrating the coarticulatory cues of /s/-retraction as

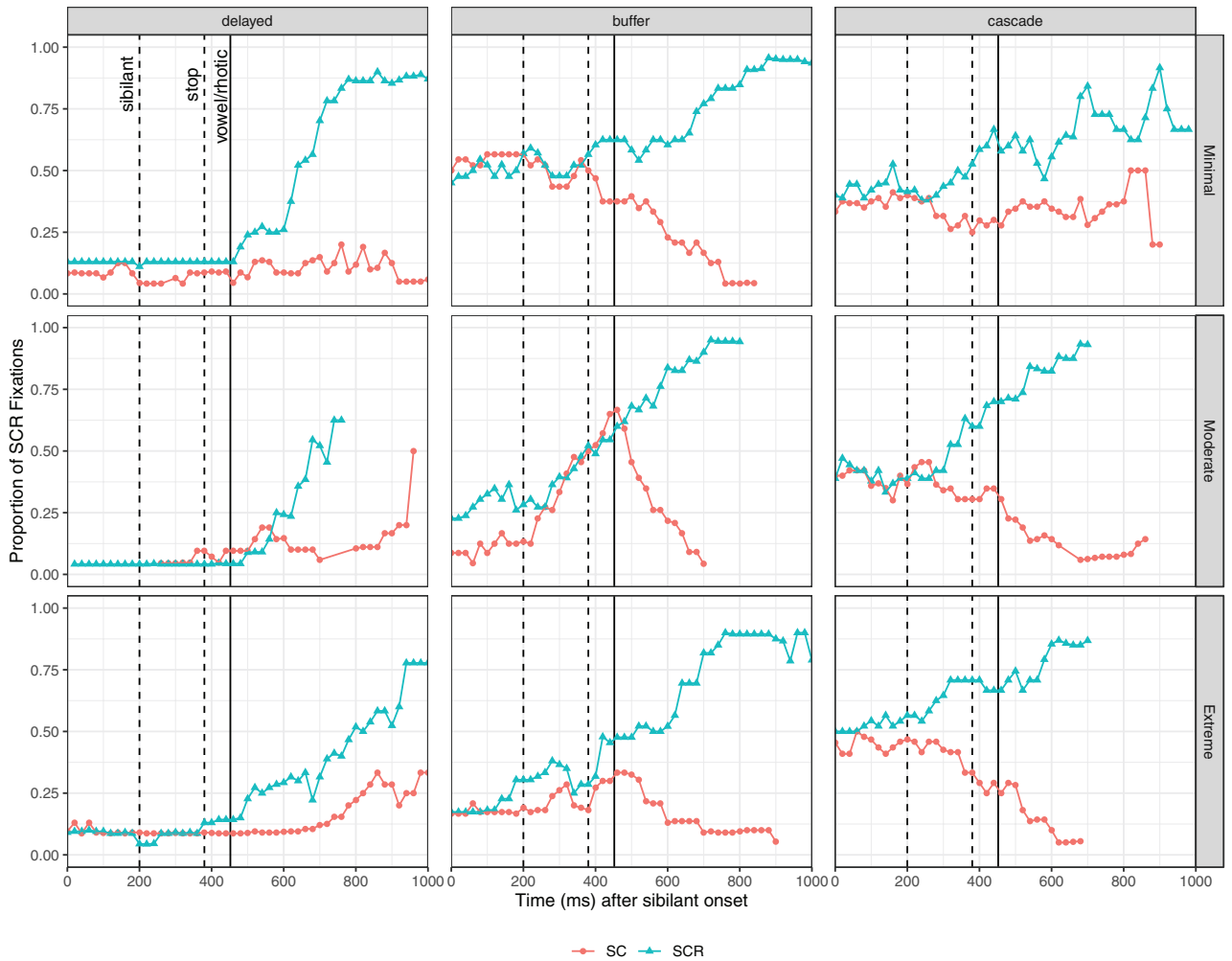


**FIGURE 5 |** Individual fixation proportion for /sCr/ clusters (y-axis) by time following the sibilant onset (x-axis, binned into 20 ms windows), cluster type (color: /sC/ = red circles, /sCr/ = teal triangles), retraction condition (rows), and pattern exhibited (columns). The vertical lines represent a 200 ms delay from the onset of the sibilant, stop, and vowel/rhotic. A fixation executed during the sibilant interval would be observed between the black dashed vertical lines.

soon as they are available to anticipate the upcoming /r/. This is not to say that all individual variation falls categorically into one of these three patterns, as intermediate strategies were observed by some participants. Rather, these nine individuals demonstrate that these three very different patterns in cue processing are utilized by participants in all three retraction conditions, suggesting that even with an abundance of cues of retraction, some individuals may still wait until the stimuli are disambiguated while other individuals will begin to inform their lexical identification with the smallest of coarticulatory cues.

## 4. DISCUSSION

The present study examined eye gaze movements to ask if listeners can use spectral information from sibilants immediately in speech processing. This study focused on two different phonological environments where spectral cues in the sibilants were doing different work: prevocalic environments, where spectral cues serve as the primary means of creating a phonemic contrast between /s/ and /ʃ/, and pre-consonantal environments, where spectral coarticulatory cues can foreshadow upcoming sounds without crossing any potential category boundaries. The results demonstrate that listeners can use spectral cues in both environments to immediately increase their consideration of the correct candidate, but more often than not listeners wait until all relevant cues have been heard.

Prevocalic /s/ and /ʃ/ are highly variable and context-dependent, meaning that no cut-and-dry category boundary can be used indiscriminately. The contrast between /s/ and /ʃ/ is made on a variety of different spectral cues and no one individual cue has been found to categorize sibilants between speakers (Jongman et al., 2000). Moreover, spectral cues on sibilants not only vary significantly in different phonological contexts, but also from speaker to speaker (Stuart-Smith, 2007). With Hypothesis

1, I asked if listeners can make immediate use of spectral information in such variable sounds in order to distinguish /s/ from /ʃ/. The results support this hypothesis and demonstrate that listeners can use the spectral cues of sibilants as they unfold in order to disambiguate phonemes, demonstrating that spectral information can be useful in even the most variable sounds.

While these findings add sibilants to a long list of sounds that listeners can begin to disambiguate before all relevant cues are available, they stand in contrast to previous work asking the same question. Galle et al. (2019) examined integration strategies for prevocalic sibilants and found that listeners appear to exhibit a buffer strategy of cue integration, waiting until the onset of the vowel before planning any gaze movements. Galle et al. (2019) explored a variety of different explanations for why sibilants appear to behave differently from other sounds, from acoustic explanations regarding the higher frequency bands occupied by fricatives to their sheer variability and unreliability. It is not immediately clear how to reconcile the present findings of a cascade strategy with the buffer strategy they observed. One possibility stems from differences in instructions: Participants in the present experiment were instructed to select the correct image as "quickly and accurately as possible," while Galle et al. (2019) "encouraged [participants] to take their time and perform accurately" (p. 12). It's possible that emphasizing speed may encourage participants to immediately integrate cues that would otherwise be stored in a buffer until additional cues become available. A second possibility comes from the experiment design: This study presents listeners with two potential candidates while Galle et al. (2019) provided four potential candidates. It's possible that when listeners know the nature of the phonological contrast between the candidates, they are more likely or more able to immediately integrate the spectral cues of that contrast, but as more candidates and contrasts are included, listeners may be more likely to hold spectral information in a buffer. Finally, and perhaps most likely, the difference may stem from differences in analysis: The present study asks whether the proportion of correct fixations improves over the course of the sibilant, while Galle et al. (2019) ask at what point the effect of the onset sibilant crosses a threshold in biasing /s/ consideration. So while Galle et al. (2019) find that listeners are relatively slower in categorizing a sibilant compared to a stop consonant, the present study finds that consideration of the correct candidate significantly improves during the sibilant itself.

With it established that listeners can immediately use the spectral cues of sibilants to discriminate phonological contrasts, the pre-consonantal analysis asks if they can use the same processing strategies for context-dependent variation in order to tease apart two lexical items that may initially be phonologically identical but phonetically distinct. With Hypothesis 2, I asked if listeners can use the coarticulatory cues of /s/-retraction as soon as they are available, such that a listener that hears a retracted /s/ may consider *string* to be a more viable candidate than *sting* even before the /r/ has been heard. The results of this study support this hypothesis, as individuals were shown to increase their consideration of the correct candidate over the course of the sibilant. Furthermore, the stronger the cues of retraction available, the greater the likelihood of considering the correct candidate. Thus, listeners not only are able immediately use the spectral cues of sibilants in order to make phonological contrasts, but also to make context-dependent predictions.

Building off Hypothesis 2, I asked in Hypothesis 3 if a retracted /s/ is a better indicator of rhotic presence than a non-retracted /s/ is of its absence. This was motivated in part by Beddor et al. (2013), who found that a nasalized vowel is a better indicator of an upcoming nasal stop than an oral vowel is for an upcoming oral stop. The results of the present study are inconclusive with respect to this hypothesis. That is, I show that participants are overall more accurate in their perception of /sC/ clusters than /sCr/ clusters, but participants are more likely to correctly look to an /sCr/ image when it is manipulated to have extreme coarticulatory cues. These findings demonstrate that listeners closely attend to different cues, but not all cues are equally helpful in every environment.

Finally, with Hypothesis 4, I again posed a follow-up to Hypothesis 2 to ask if the cues of retraction are more useful in the /str/ clusters where they are most expected than in /skr/ and /spr/ where they're less expected. While /s/-retraction has received increasing sociolinguistic and phonetic attention in recent years, little work has focused on the perception of the phenomenon in situ to ask if listeners attend to those cues. With Hypothesis 4, I ask if listeners have detailed phonological knowledge about the distribution of /s/-retraction and whether they use that knowledge in their consideration of lexical candidates in real time. The results of the present study appear to support this hypothesis as listeners exhibit increased accuracy in the consideration of /str/ clusters over the course of the sibilant compared to /spr/ and /skr/ clusters. However, it is worth noting that there is a potential confound here: not all /sCr/ clusters were manipulated to contain the same degree of retraction in the same conditions. Rather, each place series was manipulated independently relative to the model talker's baseline. Thus, alveolar /str/ clusters contain a greater proportion of /ʃ/ spectral energy than /spr/ and /skr/ clusters in each retraction condition. While this methodology maintains the natural inequalities in retraction that would be observed outside of the lab, it potentially obfuscates our understanding of the results. Is it the case that listeners show greater evidence for immediate integration of coarticulatory cues in /str/ clusters because they expect retraction in those clusters or because, like outside the lab, that is precisely where they are presented with the strongest cues of retraction?

The results of this study demonstrate that listeners can immediately integrate the spectral cues of sibilants in a laboratory setting when they know the nature of the contrast: In a *sip-ship* trial, listeners are expecting a phonological contrast between /s/ and /ʃ/ and, in a *sting-string* trial, they are anticipating or identifying whether the stimulus ultimately contains an /r/. However, it remains to be seen whether this effect can be observed outside of the lab or whether it persists in a more naturalistic task where multiple contrasts may be under consideration simultaneously. For example, if four potential candidates were provided in a trial, e.g., *sing-sting-string-shingle*, a listener is not only making a phonemic contrast between /s/ and /ʃ/ but also anticipating and identifying potential upcoming consonants. In such a scenario, a listener simply may be more likely to use a

buffer strategy. However, there may also be a false impression of buffering, if, for example, consideration of *sing, sting*, and *string* may improve but consideration of *shingle* decreases. In this hypothetical trial, looks to the correct candidate may not diverge from other potential candidates, suggesting a buffer strategy, despite the fact that the listener is actively removing other potential candidates from consideration, indicating a cascade strategy. Furthermore, while the present study focused only on the time window during which the sibilant was heard, increasing the number of potential candidates and contrasts also changes the point at which those sounds are disambiguated: Prevocalic stimuli, like *sing* and *shingle*, are disambiguated at the end of the sibilant, but pre-consonantal stimuli, like *sting* and *string* remain temporarily ambiguous. One tool we could use to tease apart these temporal differences would be to consider the integration strategies for nonce words, like *stimble-shtimble-strimble-shtrimble*. While lacking in the temporal resolution that eye tracking allows, Phillips and Resnick (2019) examined categorization of such nonce words, demonstrating that listeners on the whole are reluctant to categorize pre-consonantal onsets as /ʃ/. As listeners uphold the phonotactic restrictions of English even in the perception of nonce words, it is unlikely that nonce words would provide novel or informative evidence for cue integration strategies of pre-consonantal sibilants. Moreover, it is this phonotactic restriction on pre-consonantal sibilants that creates the space for coarticulation to vary so dramatically, giving rise to sound change emergence without endangering a phonemic contrast.

More than asking whether listeners can immediately integrate the coarticulatory cues on sibilants to aid in speech processing, this study asks whether listeners can use the variable cues of a sound change in progress. If a change is underway, it may be the case that listeners are highly variable not just in whether they attend to the cues of retraction, but also in what their acoustic expectations for /str/ clusters are or even in what their phonological representations are, i.e., /str/ vs. /ʃtr/. The present study assumes that listeners retain an underlying /s/, in part due to the phonotactic restrictions that allow even the most extreme [ʃ] to be categorized as /s/ pre-consonantally and in part due to the orthographic biases that may favor a retained /s/. Regardless of its underlying representation, /s/-retraction can help distinguish /str/ clusters from not only /st/ clusters, as examined through the present study, such that *string* and *sting* are readily disambiguated, but also /str/ clusters from /s/ onsets, such that *string* and *sing* are also disambiguated before the end of the sibilant. In its current state, where /s/ is generally intermediate between a canonical /s/ and /ʃ/, /s/-retraction is unlikely to create temporary ambiguity between /str/ clusters and /ʃ/ onsets, such that *street* and *sheet* would be initially confusable. However, it is possible, should phonological reanalysis occur or should /s/ be allophonically produced as [ʃ] in /str/ clusters, that /s/-retraction introduces a new temporary ambiguity between /str/ (or /ʃtr/) clusters and /ʃ/ onsets. This is not tested in the present experiment and the current state of /s/-retraction outside of the laboratory does not predict such a categorical [ʃ] realization, yet it remains a possibility that increased coarticulatory cues do not always disambiguate all phonological environments.

The examination of the individual listeners' results suggests that a range of different patterns were observed in each experimental condition, which demonstrate that individuals can use the cues of /s/-retraction even when they are weak. However, they need not always, as many participants show no such evidence of immediate integration. Given the nature of /s/-retraction as a change in progress, it's not clear in the present design whether listeners' unequal experiences with the change in progress can influence the robust individual variability observed. There was no effect of listener age as all participants were college-aged. Additionally, there was no effect of geographic region, which may initially be unexpected. However, /s/-retraction is a sound change noted for not being associated with any single region or demographic and has instead been referred to as a "general American innovation" (Shapiro, 1995). It is possible that regardless of how geographic region was treated, including using a rural/urban divide, geographic generalizations about the state of /s/-retraction could not capture the distribution of the change in progress. Additionally, it's possible that the geographic variation has been neutralized or diminished since all participants were members of the same community in Chicago at the time of the study and may have had similar exposure to the sound change following their formative years apart.

Furthermore, two other factors that may explain the individual variation were not included in the present design: listeners' categorical judgments and production. It is possible that if we had a means of discerning listeners' underlying representations or if we had examined at what point listeners will categorize an /str/ cluster as an /ʃtr/ cluster, that these would help predict listener variability. For instance, a listener who has an underlying /ʃtr/ cluster may immediately attend to the spectral cues as there's a phonemic contrast in play, rather than a question of coarticulation foreshadowing upcoming sounds. Speakers' own production, that is whether they produce significant retraction in /str/ clusters, may also predict their reliance on the spectral cues of retraction. We might predict that a speaker who produces more retraction may be more likely to immediately integrate the relevant cues. Alternatively, it is possible that a speaker who produces more retraction will attend only to the cues of extreme retraction (to the exclusion of moderate and minimal retraction) while a speaker who produces less retraction will attend only to the cues of minimal retraction (to the exclusion of moderate and maximal retraction). This would mirror findings from an imitation task by which only extreme retractors exhibited convergence in extreme retraction conditions, even if that meant reducing their relative degree of /s/-retraction, and only minimal retractors exhibited convergence in minimal retraction conditions, even if that meant increasing their relative degree of /s/-retraction (Phillips, 2020). At stake here is whether experience with the sound change makes a listener more sensitive to the cues across the board or whether a listener is more sensitive to cues that better align with their own speech. I leave these questions to future work and recognize that the individual variability observed here is robust even if it is not predictable.

## 5. CONCLUSION

The boundary between /s/ and /ʃ/ is anything but a clear and reliable line, clouded by mountains of ambiguity and variability. Listeners attend to the vast amount of information at their disposal to constantly shift the boundaries, whether that be because of phonological contexts, some facet of the speaker's identity, or simply as a result of the sounds they were recently exposed to Kraljic et al. (2008). This means that there is a lot of potentially conflicting information that listeners have to deal with in a short span of time. It was perhaps unsurprising that Galle et al. (2019) suggested that sibilants are possibly too variable and unreliable to be immediately integrated. Rather, listeners were thought to sit through a few milliseconds of ambiguity and wait until they have all the relevant information they need to start processing. Yet the present study finds the opposite: Despite the notable variability, or perhaps because of it, listeners are able to immediately use the cues available to them to begin lexical identification. It's worth noting that just because they can, does not mean that they must, as fixation accuracy does not cross 50% until after the vowel onset.

Moreover, the present study finds that listeners not only immediately use cues in contrasting different sibilants like /s/ and /ʃ/, but also in pre-consonantal environments where no phonological contrast between /s/ and /ʃ/ exists. In these environments, unconstrained by phonological contrasts, /s/ shows extreme coarticulatory variability, approaching the /s/-/ʃ/ boundary. This study demonstrates that listeners are astutely aware of this coarticulatory variability and use it in real-time to disambiguate words like *string* and *sting* that should be ambiguous at that point in time. Beyond demonstrating that listeners have detailed knowledge of the sound change and use that knowledge in perception, these results make interesting implications for the future of /s/-retraction as a sound change. Firstly, the results of this study demonstrate that listeners are attending to coarticulatory cues in /spr/ and /skr/ clusters despite the fact that retraction is currently much more advanced in /str/ clusters. This suggests that these environments may be the next loci for the sound change, following many other Germanic languages (Bukmaier et al., 2014). Secondly, it demonstrates that the spectral cues on /s/ serve an important role in contrasting /sC/ and /sCr/ sequences. While the results still clearly suggest that the presence or absence of /r/ is the primary disambiguating force in words like *string* and *sting*, the fact remains that listeners are carefully attending to the sibilant, in part because it temporally precedes the /r/. If the sound change continues to advance and if listeners begin to reanalyze the onset as /ʃ/, it is possible that listeners will begin to shift cue weight onto the onset sibilant until it is the primary cue in contrasting these clusters. In this scenario, the rhotic itself would eventually become redundant, which may lead to it being reduced or deleted entirely. In a possible distant future, the contrast would not be between *string* and *sting*, but *shting* and *sting*.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Chicago Social & Behavioral Sciences IRB. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

JP conceived, designed, and conducted the experiment, analyzed the data, and wrote the article.

## FUNDING

This research was supported by the National Science Foundation Doctoral Dissertation Research Improvement Grant No. BCS-1749342 and the University of Chicago-Mellon Foundation Dissertation Completion Fellowship.

## ACKNOWLEDGMENTS

Thank you to Alan C. L. Yu for his mentorship and guidance and to Diane Brentari, Susan Lin, and Jane Stuart-Smith for their valuable feedback. Thank you to Ming Xiang for the generous use of the Language Processing Laboratory facilities and equipment. Thank you to Alex Kramer for technological assistance and to Sasha Elenko and Max Fennell-Chametzky for their assistance in data collection. This article is based on Chapter 3 of the author's dissertation (Phillips, 2020).

## REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Bailey, G., Nichols, S., Turton, D., and Baranowski, M. (2022). Affrication as the cause of /s/-retraction: evidence from Manchester English. *Glossa* 7. doi: 10.16995/glossa.8026

Baker, A., Archangeli, D., and Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Lang. Variat. Change* 23, 347–374. doi: 10.1017/S0954394511000135

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., and Brasher, A. (2013). The time course of perception of coarticulation. *J. Acoust. Soc. Am.* 133, 2350–2366. doi: 10.1121/1.4794366

Bukmaier, V., Harrington, J., and Kleber, F. (2014). An analysis of post-vocalic /s-ʃ/ neutralization in Augsburg German: evidence for a gradient sound change. *Front. Psychol.* 5:828. doi: 10.3389/fpsyg.2014.00828

Connine, C. M., and Darnieder, L. M. (2009). Perceptual learning of co-articulation in speech. *J. Mem. Lang.* 61, 368–378. doi: 10.1016/j.jml.2009.07.003

Darwin, C. (2005). *Digital Mixing Script*. Brighton: University of Sussex.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.* 55, 149–179. doi: 10.1146/annurev.psych.55.090902.142028

Durian, D. (2007). "Getting stronger every day?: more on urbanization and the socio-geographic diffusion of (STR)," in *University of Pennsylvania Working Papers in Linguistics, Vol. 13*, eds S. Brody, M. Friesner, L. Mackenzie, and J. Tauberer (Philadelphia, PA), 65–79.

Elman, J., and McClelland, J. (1986). "Exploiting the lawful variability in the speech wave," in *Invariance and Variability of Speech Processes*, eds J. S. Perkell and D. Klatt (Hillsdale, NJ: Lawrence Erlbaum Associates).

Fowler, C. (1996). Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.* 99, 1730–1741. doi: 10.1121/1.415237

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *J. Phonet.* 14, 3–28. doi: 10.1016/S0095-4470(19)30607-2

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Percept. Psychophys.* 68, 161–177. doi: 10.3758/BF03193666

Galle, M. E., Klein-Packard, J., Schreiber, K., and McMurray, B. (2019). What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. *Cogn. Sci.* 43:e12700. doi: 10.1111/cogs.12700

Glain, O. (2013). *Les cas de palatalisation contemporaine (CPC) dans le monde anglophone* (Ph.D. thesis). Université Jean Moulin, Lyon, France.

Gylfadottir, D. (2015). "Shtreets of Philadelphia: an acoustic study of /str/-retraction in a naturalistic speech corpus," in *University of Pennsylvania Working Papers in Linguistics, Vol. 21* (Philadelphia, PA), 2–11.

Holt, L. L., and Kluender, K. R. (2000). General auditory processes contribute to perceptual accommodation of coarticulation. *Phonetica* 57, 170–180. doi: 10.1159/000028470

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.* 108, 1252–1263. doi: 10.1121/1.1288413

Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition* 107, 54–81. doi: 10.1016/j.cognition.2007.07.013

Lawrence, W. P. (2000). /str/ → /textipaStr/: assimilaton at a distance? *Am. Speech* 75, 82–87. doi: 10.1215/00031283-75-1-82

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, eds W. J. Hardcastle and A. Marchal (Dordrecht: Kluwer Academic Publishers), 403–439. doi: 10.1007/978-94-009-2037-8_16

Lotto, A. J., and Kluender, K. R. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Percept. Psychophys.* 60, 602–619. doi: 10.3758/BF03206049

Martin, J. G., and Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *J. Acoust. Soc. Am.* 69, 559–567. doi: 10.1121/1.385484

Matthies, M. L., Perrier, P., Perkell, J. S., and Zandipour, M. (2001). Variation in anticipatory coarticulation with changes in clarity and rate. *J. Speech Lang. Hear. Res.* 44, 340–353. doi: 10.1044/1092-4388(2001/028)

McMurray, B., Clayards, M., Tanenhaus, M. K., and Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychon. Bull. Rev.* 15, 1064–1071. doi: 10.3758/PBR.15.6.1064

Ostreicher, H., and Sharf, D. (1976). Effects of coarticulation on the identification of deleted consonant and vowel sounds. *J. Phonet.* 4, 285–301. doi: 10.1016/S0095-4470(19)31256-2

Phillips, J. B. (2020). *Sibilant categorization, convergence, and change: the case of /s/-retraction in American English* (Ph.D. thesis). University of Chicago, Chicago, IL, United States.

Phillips, J. B., and Resnick, P. (2019). Masculine toughness and the categorical perception of onset sibilant clusters. *J. Acoust. Soc. Am.* 145:EL574. doi: 10.1121/1.5113566

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Scarborough, R. (2004). *Coarticulation and the structure of the lexicon* (Ph.D. thesis). University of California, Los Angeles, Los Angeles, CA, United States.

Shapiro, M. (1995). A case of distant assimilation: /str/ → /ʃtr/. *Am. Speech* 70, 101–107. doi: 10.2307/455876

Smith, B. J., Mielke, J., Magloughlin, L., and Wilbanks, E. (2019). Sound change and coarticulatory variability involving English /ɹ/. *Glossa* 4:63. doi: 10.5334/gjgl.650

Stevens, K., and Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *J. Phonet.* 38, 10–19. doi: 10.1016/j.wocn.2008.10.004

Stevens, M., and Harrington, J. (2016). The phonetic origins of /s/-retraction: acoustic and perceptual evidence from Australian English. *J. Phonet.* 58, 118–134. doi: 10.1016/j.wocn.2016.08.003

Stilp, C. (2019). Acoustic context effects in speech perception. *WIREs Cogn. Sci.* 11:e1517. doi: 10.1002/wcs.1517

Stuart-Smith, J. (2007). "Empirical evidence for gendered speech production: /s/ in Glaswegian," in *Laboratory Phonology, Vol. 9*, eds J. Cole and J. I. Hualde (New York, NY: Mouton de Gruyter), 65–86.

Viswanathan, N., Magnuson, J. S., and Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *J. Exp. Psychol.* 36, 1005–1015. doi: 10.1037/a0018391

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Percept. Psychophys.* 50, 351–360. doi: 10.3758/BF03212227

Wilbanks, E. (2017). "Social and structural constraints on a phonetically motivated change in progress: (STR) retraction," in *Working Papers in Linguistics, Vol. 23* (Philadelphia, PA). doi: 10.5070/P7121040720

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.