Resolving Competing Predictions in Speech: How Qualitatively Different Cues and Cue Reliability Contribute to Phoneme Identification

Anne Marie Crinnion[1], Sahil Luthra[2], Phoebe Gaston[1], and James S. Magnuson[1,3,4]

[1] University of Connecticut

[2] Carnegie Mellon University

[3] BCBL, Basque Center on Cognition, Brain and Language

[4] Ikerbasque, Basque Foundation for Science

Author Note

This research was supported by National Science Foundation (NSF) grants NRT 1747486 and PAC 2043903 (PI JSM). Phoebe Gaston was supported by NSF grant CRCNS 2207770. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Anne Marie Crinnion and Phoebe Gaston were supported by NIH T32 DC017703. Sahil Luthra was supported by NIH F32 DC020625. Anne Marie Crinnion was supported by F31 DC021372. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported in part by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through project PID2020-119131GB-I00 (BLIS). Preliminary results from this work were presented at the 62nd Annual Meeting of the Psychonomic Society. We thank Rachel Theodore, Emily Myers, and Jay Rueckl for helpful discussion of this work. PG is now at McMaster University.

Corresponding Author

Anne Marie Crinnion

anne.crinnion@uconn.edu

Abstract

Listeners have many sources of information available in interpreting speech. Numerous theoretical frameworks and paradigms have established *that* various constraints impact the processing of speech sounds, but it remains unclear *how* listeners might simultaneously consider multiple cues, especially those that differ qualitatively (i.e., with respect to timing and/or modality) or quantitatively (i.e., with respect to cue reliability). Here, we establish that cross-modal identity priming can influence the interpretation of ambiguous phonemes (Exp. 1, N = 40) and show that two qualitatively distinct cues — namely, cross-modal identity priming and auditory coarticulatory context — have additive effects on phoneme identification (Exp. 2, N = 40). However, we find no effect of quantitative variation in a cue — specifically, changes in the reliability of the priming cue did not influence phoneme identification (Exp. 3a, N = 40; Exp. 3b, N = 40). Overall, we find that qualitatively distinct cues can additively influence phoneme identification. While many existing theoretical frameworks address constraint integration to some degree, our results provide a step towards understanding *how* information that differs in both timing and modality is integrated in online speech perception.

Introduction

When trying to understand speech, listeners can use many different types of information beyond just the acoustic signal itself. In fact, speech frequently occurs in complex auditory environments (e.g., multi-talker environments; Cherry, 1953; Schneider et al., 2007) and often is not carefully articulated (for review, see Smiljanić & Bradlow, 2009), both of which increase ambiguity in the signal. The fact that speech perception is typically robust despite this ambiguity hints at the fact that in a given listening situation, there are myriad constraints that help the listener understand what is being said. Knowledge about who is talking (Kleinschmidt, 2019; Kraljic & Samuel, 2007; Nygaard & Pisoni, 1998), the topic being discussed (Borsky et al., 1998; Broderick et al., 2019; Hutchinson, 1989; Liberman, 1963), and how various sounds tend to pattern in one's language (Frisch et al., 2000; Leonard et al., 2015; Mersad & Nazzi, 2011) are just a few examples of co-occurring constraints that may be available for processing. Language scientists have spent decades investigating how people make use of different types of information to process language, with a major aim of understanding what constraints are relevant in different contexts and how constraints might be combined (for an overview see McRae & Matsuki, 2013). While theories of constraint satisfaction in language processing have been developed at various levels (MacDonald et al., 1994; Trueswell & Tanenhaus, 1994) and have been extended to domain-specific computational models (e.g., the TRACE model of speech perception and spoken word recognition; McClelland & Elman, 1986; SOPARSE, a model of sentence processing; Tabor & Hutchins, 2004), exactly how *qualitatively distinct constraints* are combined remains outside of the scope of these theories.

Bayesian approaches (e.g., Kleinschmidt & Jaeger, 2015) provide a computational framework in which expectations regarding the relative weight and reliability of different possible constraints (*priors*, such as knowledge of a talker's typical productions, or of contextual information) can be integrated according to principles following from Bayes' theorem to integrate them optimally. While this framework establishes a mathematical model of how an ideal observer should combine cues (in terms of Marr's [1982] levels of information processing theories, it is a theory of the computations required), it does not provide an algorithmic mechanism for this cue integration.

We are concerned with the algorithmic level — that is, when and how various types of constraints are processed during online processing, in the context of real-time pressures on listeners. In the TRACE model of speech perception, for example, different types of constraints can be simulated (i.e., lexical-level constraints on phoneme ambiguities, or acoustic-level constraints, such as coarticulatory effects; Elman & McClelland, 1988). The relative weighting of two constraints, or how they are actually combined, is not inherent in the model architecture. For example, in order to simulate effects akin to preceding (lexical or semantic) context, TRACE has a priming mechanism that allows the experimenter to boost the resting activation of specific words. However, it is not immediately clear how priming in TRACE could be parameterized to appropriately model over-time influences or interactions with other cues. Doing so would require empirical data that clarify how effects of prior context interact with other, qualitatively distinct cues to influence speech processing.

The goal of this work, then, is to understand how both *qualitative* and *quantitative* factors influence the use of multiple constraints in speech perception. We ask how constraints that differ *qualitatively* (in terms of timing and modality) and *quantitatively* (in terms of reliability) affect processing of ambiguities in speech. In the General Discussion, we return to the challenges multiple cue integration poses for models of human spoken word recognition.

Qualitative variation in constraints

Information that can influence the identification of speech sounds can stem from a variety of sources. Some are related to variations in the actual acoustic signal. Others are based on linguistic knowledge (e.g., constraints implied by the words the listener knows, or expectations based on syntax or semantics). Many different kinds of constraints may be available in real-world contexts. Imagine the following scenario: an individual walks into the kitchen, where their roommate is holding a shopping bag and says, "*The chiropractor dealt with my ba[?]*", where the final word contains a segment ambiguous between /g/ and /k/ (*bag* or *back*). This scenario contains cues in different modalities, occurring at different time points. The visual information (the roommate holding a bag) occurs early on and has the possibility to suggest that the roommate said "bag". There is also semantic context from the word *chiropractor* that might bias the listener towards hearing "back". This cue occurs within the speech modality but still precedes the ambiguous segment by hundreds of milliseconds. There may also be constraining acoustic context adjacent to the point of ambiguity (e.g., the duration of the preceding vowel, where a longer vowel would be more consistent with the voiced alternative, /g/; Denes, 1955). There can be, then, many sources of constraint in different modalities and with different temporal relations to the point of ambiguity, and these can also vary in how informative or reliable they will be. How listeners use qualitatively different cues when processing speech, particularly when these sources of information may conflict, is not fully specified by psycholinguistic theories.

Much of the work on ambiguity resolution in speech perception looks at the effect of just one cue on another (often, how the perception of a phonetic continuum is influenced by the presence of a particular cue). Consider the well-known Ganong effect, which isolates lexical knowledge as a potential influence on ambiguity resolution (Ganong, 1980). For example, if an ambiguous /s/-/ʃ/ token is embedded in a *sang-*shang* continuum, listeners are more likely to identify ambiguous continuum steps as /s/, since only *sang* is a word. Semantic information can likewise influence ambiguity resolution. For example, Getz and Toscano (2019) found that semantic context can influence perception of targets from minimal pairs that differ in voicing (e.g., seeing the visual prompt AMUSEMENT before an auditory token ambiguous between *bark* and *park* leads to more /p/ responses).

Other studies have introduced multiple constraints on phonetic interpretation (e.g., trading relations; Repp, 1982). Typically, however, these studies look at use of multiple acoustic-phonetic cues. For example, many studies show that listeners seem to integrate across acoustic-phonetic cues (e.g., voice-onset times, vowel durations, preceding rate) and use these cues when they become available, reflecting a process of continuous integration (McMurray et al., 2008; Reinsch & Sjerps, 2013; Toscano & McMurray, 2015). This rapid integration of cues has also been shown at the lexical level, in a Ganong paradigm; listeners use lexical information very early (Kingston et al., 2016).

Recent work from Kaufeld et al. (2020) looked at the influence of cues across levels of language processing (i.e., not just cues stemming from the acoustic signal). They studied both syntactic and acoustic constraints on phoneme identification and found additive effects of these sources of information in how listeners resolve ambiguities in the speech signal. In their study, listeners were influenced by both speaking rate (acoustic-phonetic) and morphosyntactic gender information (lexical/semantic) in interpreting ambiguous words in sentences. A similar examination of multiple sources of information in the visual word recognition literature comes from studies of how word frequency and masked repetition priming interact (Balota & Spieler, 1999; Becker, 1979; Connine et al., 1990; Forster & Davis, 1984; Holcomb & Grainger, 2006). Kinoshita (2006) showed both additive and interactive effects of these two sources of information, depending on how familiar the presented words were. When low frequency words were familiar, priming effects were greater for lower frequency words than higher frequency words. However, when low frequency words included familiar and unfamiliar items, priming effects did not differ based on lower or higher frequency. While familiarity is not of interest in the current work, it is interesting to note that under certain conditions, two sources of information might both appear to have influence, while under other conditions, the presence of one cue might eliminate, diminish, or amplify the influence of another to different degrees, at different levels of the first cue.

Recent work from Lai and colleagues (2022) has directly tested how two qualitatively distinct constraints influence spoken word recognition. Using a lexical decision task, they assessed how listeners used both coarticulatory information (stemming from rounded vowels following a sibilant, biasing listeners towards hearing /s/) and lexical information (*assume* vs. **ashume*). Consistent with findings from Kaufeld and colleagues (2020), Lai and colleagues (2022) found additive effects overall of these sources of information. Additionally, like Kaufeld et al. (2020), they also found differences at the individual level in the types of information that listeners tended to use. Namely, listeners who used lexical information more, for example, tended to use coarticulatory information less.

We note, however, that Lai and colleagues used word-nonword continua (e.g., *assume/*ashume*) for their stimuli. Given the strength of the Ganong effect, it is possible that their paradigm might underestimate the influence of coarticulatory information. Hence, in situations where listeners only hear words, it is possible that they might rely more on coarticulatory information. Additionally, we note that the lexical information comes after the point of ambiguity, which could influence the time course of lexical influences on phonetic processing and potentially modulate the strength of coarticulatory constraints.

While these studies provide steps towards understanding how different constraints may operate simultaneously, we must also consider the fact that quantitative differences, such as how reliable different constraints are, may influence how much a given constraint is used.

Quantitative variation in constraints

An important factor that might impact how listeners balance constraints from two sources is how reliable each source of information is. Work from Bushong and Jaeger (2019) suggests that in laboratory settings when there are unnatural correspondences between acoustic and contextual cues, listeners tend to discount contextual information. For instance, in naturalistic settings, /d/-like voice onset times (VOTs) occur more often in sentence contexts containing /d/-initial words, but these cue correspondences are often violated in experimental settings, where listeners may be presented with a high proportion of inconsistent cues (e.g., /t/-like VOTs in semantic contexts consistent with /d/).

Bushong and Jaeger presented listeners with sentences that varied in semantic context and VOT of the target word (example sentences from their paradigm include: (A) *When the [?]ent in the forest was well camouflaged, we began our hike*, and (B) *When the [?]ent in the fender was well camouflaged, we sold the car*). The ambiguous token, *[?]ent*, varied across six VOT values (from most /t/-like to most /d/-like). In one condition, which had "high conflict" between lexical expectations and acoustic cues, the six possible VOT values were evenly distributed across semantic contexts. Note that this condition mirrors the setup of most laboratory studies, where the goal is to obtain an equal number of observations per cell. In this high conflict condition, listeners were equally likely to hear a semantically consistent sentence, containing a phrase such as “dent in the fender”, as they were to hear a semantically inconsistent sentence, containing a phrase such as “tent in the fender”. In their "low conflict" condition, VOT values occurred in more natural proportions with typical lexical-VOT distributions (with endpoint tokens only presented with their expected lexical context and gradually decreasing proportions of consistency up to the completely ambiguous tokens, which were presented equally often with both lexical contexts). For example, in the low conflict condition, listeners heard a higher proportion of /t/-like VOTs in the semantically consistent forest sentence context (“tent in the forest”). Listeners used lexical context more in the low conflict condition, suggesting that listeners may be sensitive to the distributions of (sometimes competing) cues.

Similar sensitivity to the distribution of cues has been shown by Giovannone and Theodore (2021) in a Ganong (1980) paradigm, where listeners heard tokens along a **giss-kiss* continuum (where lexical context biases towards /k/) and along a *gift-*kift* continuum. When the degree of conflict between lexical and phonetic cues was reduced (i.e., in a low conflict condition, listeners heard a higher proportion of /g/-like VOTs in a *gift-*kift* continuum), listeners also seemed to rely more on lexical information.

Additionally, at the acoustic level, listeners may down-weight the importance of an acoustic cue depending on whether it agrees with other disambiguating information, whether acoustic or lexical (Idemaru & Holt, 2011; Zhang et al., 2021). For example, Idemaru and Holt (2011) presented listeners with pairs of cues (e.g., VOT and F0, which are both cues to voicing) that were either consistent with typical correlations between the cues, or reversed (e.g., an F0 value that typically occurs for a voiced token paired with a VOT for a voiceless token). In multiple pairings of cues, they found that when listeners were tested on ambiguous tokens where only one cue (F0) could guide their phoneme decision, they relied less on that cue (F0) when it had previously been paired with an atypical VOT value. This suggests that listeners track distributions of cues and adjust their reliance on them in accordance with recent experience.

In sum, listeners can alter their reliance on specific cues in the presence of input that either mirrors or violates naturally occurring cue correspondences. In the context of qualitatively distinct cues, it remains unknown exactly how listeners might use information about the reliability of one cue to guide their relative cue use. In other words, it is unknown if the additive or interactive effects of distinct cues changes when the reliability of one source of information changes. The current study aims to address the question of how listeners use two qualitatively different cues (cross-modal identity priming and coarticulatory context, which differ in content, modality, and temporal proximity to ambiguities in our materials) to interpret ambiguities in the speech signal, and whether listeners are sensitive to the reliability of certain cues (i.e., whether they differentially use certain cues based on their reliability).

An approach to studying constraint integration

Our aim with the current experiments is to examine how listeners use two constraints, coarticulatory context and cross-modal identity priming (henceforth referred to as visual priming), that differ saliently in modality and in temporal proximity to the point of ambiguity. Previous work has shown that coarticulatory context alone can influence the identification of ambiguous phonemes (e.g., Luthra, Peraza-Santiago, Beeson, et al., 2021), and in Experiment 1, we test whether visual priming alone can also guide the interpretation of ambiguous phonemes. In Experiment 2, we examine how these two cues might be used when both are available.

Finally, in Experiment 3, we examine whether manipulating reliability of the visual prime leads to different use of that cue. To set the stage, we conclude this section with a review of the two constraints we will manipulate in the experiments.

The first constraint will be a written word-form with potential to influence processing of a corresponding spoken word through visual priming (Blank & Davis, 2016; Sohoglu et al., 2014). Previous studies suggest that such primes influence how speech is perceived. For instance, for both degraded (vocoded) and relatively clear speech, Blank and Davis (2016) found that participants had greater accuracy reporting what they had heard when auditory stimuli had been preceded by a visual identity prime (e.g., written SHAME before degraded acoustic token *shame*), as compared to a neutral prime (e.g., written ######## before acoustic token *shame*). However, to our knowledge, previous work has not examined whether visual priming can shift the identification of ambiguous phonemes.

The second constraint will come from coarticulatory context. Effects of coarticulatory context can be seen in a paradigm known as *compensation for coarticulation* (CfC; Mann, 1980; Mann & Repp, 1981; Repp & Mann, 1981, 1982; Viswanathan et al., 2010). When speakers produce a sound with a posterior place of articulation (PoA), such as /k/, and then produce a sound with an anterior PoA (e.g., /s/) (or the other way around), speakers may not reach the typical PoA on the second sound and subsequently produce a more ambiguous speech sound. Hence, if listeners hear a token like *maniac* (with word-final /k/ and therefore posterior PoA) followed by an ambiguous *same-shame* token, they will be more likely to interpret the ambiguous token as “same” (with anterior PoA), as though they are compensating for acoustic contingencies that follow from coarticulation. Though there are alternative explanations for CfC effects that appeal to acoustic differences rather than articulatory differences (Diehl et al., 2004;

Holt & Lotto, 2008; but see Viswanathan et al., 2010), for the purposes of this investigation, it only matters that these effects exist as another instance of context influencing interpretation of speech and that these effects differ from the effects of visual priming with regards to timing and modality.

Thus, in a series of four pre-registered experiments (see preregistrations at https://osf.io/6kmub), we will compare the impact of two competing constraints on phoneme identification. These constraints vary qualitatively in modality (visual vs. auditory) and in their temporal relation to the point of ambiguity in the speech signal (visual primes will occur more than one second before the point of ambiguity, while coarticulatory context will immediately precede the point of ambiguity). In line with prior work looking at integration of various acoustic-phonetic sources of information (e.g., McMurray et al., 2008; Toscano & McMurray, 2015), it is possible that the constraints we consider influence processing from the moment they are available. However, it is also possible that one constraint dominates. Additionally, because of the temporal order inherent in the presentation of these two constraints, it is possible that the coarticulatory information only exerts an influence when it is in conflict with the prime (which may already maximally activate the target lexical or phonetic item).

We will also examine whether more reliable information coming from the visual primes (i.e., including a greater proportion of trials where the prime matches the auditory target) leads to greater use of the prime (as in Bushong & Jaeger, 2019 or Giovannone & Theodore, 2021). How listeners use qualitatively distinct constraints with varying degrees of reliability will inform theories of language processing, and provide a foundation for extending algorithmic accounts of speech processing to account for the potentially simultaneous influence of cues that are qualitatively distinct in modality and timing.

Experiment 1

In Experiment 1, we examine how visual identity priming influences identification of ambiguous word-word minimal pairs. In this study, listeners made an 's'-'sh' judgment for spoken continua created from minimal pairs like *same-shame* that were preceded by visual primes that were neutral ("########") or matched one endpoint ("SAME" or "SHAME"). Visual identity priming has been shown to influence identification of noise-vocoded speech (Blank & Davis, 2016; Sohoglu et al., 2014). However, it remains unknown whether such priming can influence perception of ambiguous tokens (such as tokens along a *same-shame* continuum). Because visual *semantic* priming has been found to influence perception of word-word pairs (Getz & Toscano, 2019), we hypothesize that identity priming should influence phoneme identification. Establishing whether visual priming can influence perception of an acoustic-phonetic continuum is a prerequisite to our goal of pitting qualitatively distinct constraints against one another in Experiment 2.

Methods

Materials

We used materials developed by Luthra, Peraza-Santiago, Beeson, et al. (2021). Luthra, Peraza-Santiago, Beeson, et al. identified context items and target pairs that elicit robust compensation for coarticulation (CfC; necessary for Experiments 2 and 3), and we used the target items (and in Experiments 2 and 3, the context items) that they established can drive CfC. We included five /s/-/ʃ/ minimal pairs that were shown to exhibit CfC effects in the pilot from Luthra, Peraza-Santiago, Beeson, et al (2021). These pairs were: *same-shame, sell-shell, sign-*

shine, sip-ship, and *sort-short*. Each pair consisted of five audio stimuli identified by Luthra, Peraza-Santiago, Beeson, et al. (2021): the most ambiguous step (proportion of /s/ responses across five pairs = 0.47) and two steps on each side of that maximally ambiguous step (where the most s-like step had a mean s-rate of 0.92 across five pairs and the most ʃ-like step had a mean s-rate of 0.04 across five pairs). For each pair, we used three written primes, with one matching each end of the target continua and one that was neutral (e.g., SIP, SHIP, and ########). In Experiments 2 and 3, we include coarticulatory context items before the /s/-/ʃ/ ambiguity. To keep the timing identical in this experiment, we inserted silent pauses, matched to the durations of the appropriate context items, between the presentation of the prime and the onset of the critical auditory target (Figure 1). The four context items were *isolate* (846 ms), *maniac* (785 ms), *pocketful* (765 ms), and *questionnaire* (1046 ms).

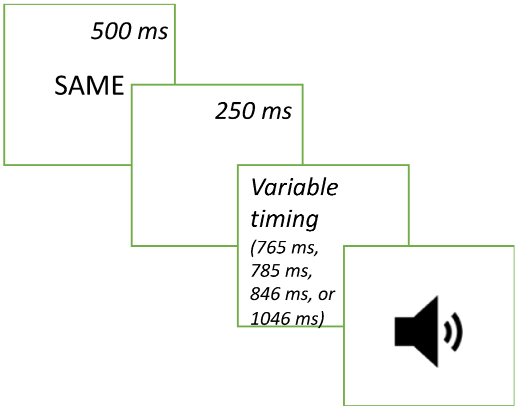

Figure 1. Schematic of a trial for Experiment 1.

Participants

We collected data from 68 participants in order to achieve our pre-registered target sample size of 40 participants (15 female, 24 male, 1 other/decline to state; age range: 19-33 years; mean age: 27 years) after applying pre-registered exclusionary criteria (described below). To determine the appropriate sample size for our experiments, we considered relevant studies examining identity priming (Blank & Davis, 2016; n = 20 for 90% power) and compensation for coarticulation (Luthra, Peraza-Santiago, Beeson, et al., 2021; n = 15 to achieve 90% power). However, a sample size sufficient for 90% power in previous studies might not be sufficient when combining constraints and examining interactions (as we do in Experiments 2 and 3). As such, we took a conservative approach and doubled the larger of the sample sizes, leading to a target sample of 40 participants per experiment.

Experimental sessions took approximately 60 minutes. Participants were paid $12, consistent with Connecticut minimum wage ($12/hr at the time of data collection). Only participants who were 18-34, native speakers of North American English, and who reported normal/corrected-to-normal vision and normal hearing were recruited for this study.

After data collection, we applied our pre-registered exclusionary criteria to exclude participants (a) for not reaching at least 80% accuracy for the clear endpoint stimuli with neutral primes, in line with conventions used in Luthra, Peraza-Santiago, Beeson, and colleagues (2021), (b) for failing to respond in more than 10% of trials (with a 6 second trial timeout), (c) for failing our headphone check (described below) more than once, or (d) for not reaching at least 80% accuracy on reporting written primes (see Procedure below).

Procedure

The experiment was implemented in Gorilla (www.gorilla.sc; Anwyl-Irvine et al., 2020) and participants were recruited through Prolific (www.prolific.co). All procedures were approved by the University of Connecticut’s Institutional Review Board (IRB). Participants provided informed consent and filled out demographic information before the main task. Participants then completed a headphone screening that required them to identify the quietest tone among a series of three tones, a task that is designed to be difficult to pass without headphones due to phase cancelation (Woods et al., 2017). If a participant failed the screening twice, we excluded their data (per our pre-registered exclusion criteria), but they still received compensation as described above.

Trials consisted of a printed prime word presented in capital letters (in Open Sans font) for 500 ms, followed by a brief pause (250 ms), a silent gap corresponding to the duration of a context item (see Materials for timing details), and an auditory target (Figure 1). Participants responded as to whether they thought the target started with an ‘s’ sound or an ‘sh’ sound by pressing the appropriate button (F or J; assignment of ‘s’ and ‘sh’ to F or J keys was counterbalanced across participants). To ensure that participants were paying attention to the written prime, on a subset of trials, participants were only presented with a written prime and asked to type that prime in a response box. Participants completed two blocks of trials, each consisting of 300 experimental trials (including all combinations of 5 target continuum steps, 5 target pairs, 3 written primes, and 4 gap durations) and 60 prime-only trials. Trial order was completely randomized within each block. Before the main blocks of the experiment, participants completed 12 practice trials with a different target continuum (*daze-gaze*). The experiment took about 60 minutes to complete.

Analyses

Pre-registered mixed-effects logistic regression models were run to predict the proportion of front PoA (i.e., /s/) responses, using the *mixed* function in the R (R Core Team, 2021) package *afex* (Singmann et al., 2015), which reports results in ANOVA-like formats and is a wrapper for the *glmer* function in the *lme4* package (Bates et al., 2015). Our model included fixed effects of Prime (front-consistent [e.g., SIP], back-consistent [e.g., SHIP], or neutral [XXXXXXXX]; sum-coded) and Step (which ranged from -2 to +2) and their interaction. Our model also included by-subject and by-target-pair random slopes for Prime and Step and their interaction, as well as by-subject and by-target-pair random intercepts, without correlation between random slopes and intercepts. Following best practices outlined by Matuschek et al. (2017), we selected this random effects structure by starting with the maximal model for our data and then using the *anova* function to test for differences between models with successively simpler random effects structures (first removing correlations between random slopes and intercepts and then removing by-item random effects) to arrive at the simplest model that does not significantly reduce fit. To investigate pairwise comparisons within the model, in exploratory analyses, we followed up on significant effects in the model using the *emmeans* package (Lenth, 2022), adjusting for multiple comparisons using the multivariate t-distribution. Details and results from pre-registered analyses of reaction time data can be found in the supplementary materials.

Results

Figure 2 shows that participants made more front-PoA responses when they had seen a front-PoA prime (e.g., SAME) as compared to either a neutral prime (e.g., ########) or a back-PoA prime (e.g., SHAME). More specifically, across all steps, participants made a front-PoA response 47% of the time after a front-PoA prime, 42% of the time after a neutral prime, and 40% of the time after a back-PoA prime.

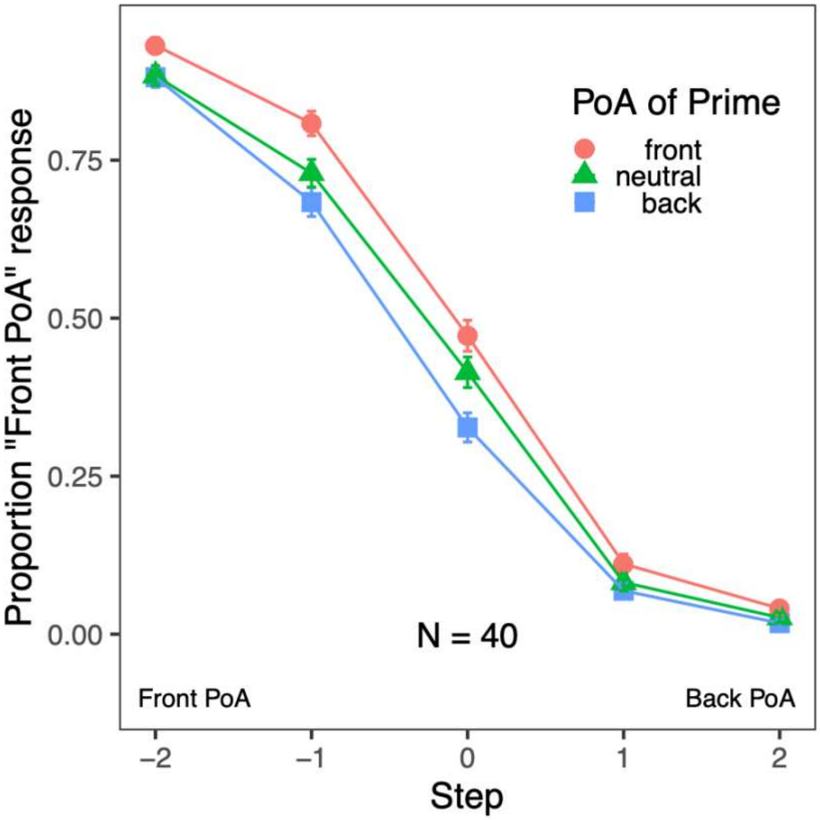

Figure 2. Responses from Experiment 1 showing decisions on the first phoneme of a target continuum (all continua were between /s/ and /ʃ/, e.g., *same-shame*) after a written-word prime. The x-axis shows continuum step, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows the proportion of front PoA (i.e., /s/) responses, with colors and shapes indicating prime type. Error bars represent 95% confidence intervals.

Our mixed effects logistic regression model revealed a significant effect of Prime (χ^2 = 16.13, p < .001), indicating that participants' responses were influenced by the written prime. After participants saw a prime beginning with a front PoA (i.e., SAME), they were more likely to make an /s/ response, which was the expected direction of this effect. The model also revealed a significant effect of Step (χ^2 = 19.20, p < .001), indicating that participants made more front-PoA responses for more front-PoA steps. The interaction between Prime and Step was not significant ($\chi^2 = 5.36$, $p = .07$).

We conducted follow-up tests to analyze pairwise comparisons for Prime, correcting for multiple comparisons as described above. There were more front-PoA responses for a front Prime than for a neutral Prime (contrast estimate: .399, *z-ratio* = 4.171, $p < .001$), and more front-PoA responses for a front Prime than for a back Prime (contrast estimate: .772, *z-ratio* = 6.040, $p < .001$). Likewise, there were also fewer front-PoA responses for a back Prime than for a neutral Prime (contrast estimate: .373, *z-ratio* = 4.272, $p < .001$).

Discussion

Overall, these findings demonstrate that visual priming influences identification of ambiguous phonemes, both influencing trial-level interpretation of the stimulus and promoting faster response times. To our knowledge, this is the first demonstration of the influence of visual priming on such a task, though as we noted above, semantic priming has been shown to influence identification of ambiguous phonemes (Getz & Toscano, 2019). In the General Discussion, we consider broader implications and potential extensions of this finding. Most importantly, however, we note that the demonstration that visual identity primes influence identification of ambiguous phonemes will allow us to examine in Experiment 2 how priming does (or does not) influence speech processing when another qualitatively different cue is present: coarticulatory context.

Experiment 2

In Experiment 2, we investigate how listeners reconcile potentially concordant or conflicting cues that differ in both modality and timing: a visual prime and coarticulatory context. We established in Experiment 1 that presenting a visual prime such as SHAME before an ambiguous *same-shame* token makes listeners more likely to identify ambiguous tokens as "shame". Prior research shows that presenting an auditory token of *isolate* (ending with an anterior PoA) before an ambiguous *same-shame* token will likewise lead listeners to be more likely to report hearing "shame" (starting with a posterior PoA), due to the CfC (Compensation for Coarticulation) effect introduced above (Mann, 1980; Mann & Repp, 1981; Luthra, Peraza-Santiago, Beeson, et al., 2021; Repp & Mann, 1981, 1982). Of interest in the current work is how listeners make use of two different sources of information, particularly when they are in conflict.

By presenting listeners with both types of information (e.g., visually priming SAME before the audio *isolate,* before an ambiguous *same-shame* token – where the two constraints make opposite predictions [since CfC based on the front-PoA at the offset of *isolate* would be consistent with *shame*]), we test how both constraints influence speech perception. If listeners are more sensitive to the earliest information available, then we might expect listeners to rely more on the prime. If, however, listeners are more sensitive to within-modality information, we might expect listeners to rely more on the auditory context that immediately precedes the target. We also might expect to observe interactive effects, such that the presence of one source of information changes the effect of the other source of information. Another possibility (consistent with prior research, e.g., Kaufeld et al., 2020; Lai et al., 2022) is that we might observe additive effects of the two constraints. Note that here, we refer to interaction in the statistical sense, with relevant implications for underlying cognitive mechanisms.

We note that while Lai and colleagues found additive effects of coarticulatory information and lexical information, there were key differences with our work. First, their constraints (coarticulatory information and lexical information) occurred within the same modality, which might better facilitate additive processing. Our work asks how constraints that differ in modality and timing constrain speech processing. Second, they used word-nonword continua, which may have magnified lexical effects, since typical processing does not involve nonwords. Third, lexical information occurred only after the point of ambiguity. In contrast, we present both lexical (visual prime) and coarticulatory information before the auditory target (which comes only from word-word continua).

Methods

Materials

Stimuli consisted of the same primes and target pairs from Experiment 1. While in Experiment 1 we inserted silent gaps matched to the durations of the context items from Luthra, Peraza-Santiago, Beeson, et al. (2021), in Experiment 2 we used the actual audio tokens of those four context items: *isolate*, *maniac*, *pocketful*, and *questionnaire.* Two of these auditory context items (*isolate* and *pocketful*) end in a front PoA and two (*maniac*, and *questionnaire*) end in a back PoA. Recall that presenting auditory context items that end with a front PoA (*isolate* and *pocketful*) before hearing an ambiguous same-shame token should lead listeners to hear *shame* (starting with a back PoA), due to CfC (Mann, 1980; Mann & Repp, 1981; Repp & Mann, 1981, 1982).

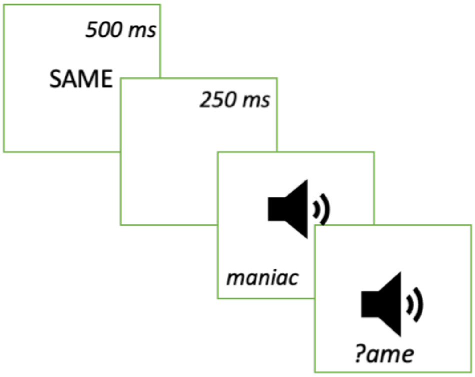

Figure 3. Schematic of a trial for Experiment 2.

Participants

We collected data from 63 participants in order to achieve our pre-registered target sample size of 40 participants (15 female, 25 male; age range: 18-34 years, mean age: 28 years) after applying pre-registered exclusionary criteria.

Experimental sessions took approximately 60 minutes. Participants were paid $12 for their participation, consistent with Connecticut minimum wage ($12/hr at the time). We used the same pre-registered recruitment and exclusionary criteria as outlined in Experiment 1.

Procedure

All procedures were the same as in Experiment 1, except that auditory context items were presented in place of the corresponding silent gaps. Thus, trials consisted of a printed prime word for 500 ms, followed by a brief pause (250 ms), an auditory context item, and an auditory target (Figure 3). As in Experiment 1, participants completed 12 practice trials that used different context items and a different target continuum (context: *catalog*; continuum: *lip*-*rip*).

Analyses

Analyses followed a similar structure to Experiment 1. For our analysis of participants' responses (i.e., whether they indicated the target began with a /s/ or /ʃ/), the logistic mixed effects model included fixed effects of Prime (front, back, and neutral; sum-coded), Context (front and back; sum-coded) and Step (which ranged from -2 to +2), and their three-way interaction (as well as all the lower-level interactions). This model also included by-subject and by-target-pair random slopes for Prime, Context, Step, and their three-way interaction (as well as all the lower-level interactions), as well as by-subject and by-target-pair random intercepts, with no correlations between random slopes and intercepts. We arrived at this random effects structure using the same model selection criteria as in Experiment 1. Pairwise comparisons were investigated following the same approach as in Experiment 1. As for Experiment 1, results from pre-registered analyses of reaction time data can be found in the supplementary materials.

Results

Responses

As shown in Figure 4, participants were influenced by both the visual prime and the auditory context items. As predicted, across all steps, participants made more front-PoA responses when they heard an auditory context item that ended in a back-PoA (overall 47%) than when the context item ended in a front-PoA (overall 38%). In other words, when participants heard an auditory context such as *isolate* (which has a front PoA), they were more likely to report that the auditory target began with /ʃ/ (which has a back PoA), which demonstrates the expected CfC effect. Likewise, they made more front-PoA responses when they saw a front prime (overall 47%) than when they saw a neutral prime (overall 42%) or a back prime (overall 39%). Breakdowns of the proportion of front responses by condition can be found in Table 1.

Table 1.

Proportions of front-PoA responses based on condition

Context	Prime			
	Front	Neutral	Back	Mean
Back	0.52	0.46	0.44	0.47
Front	0.42	0.37	0.34	0.38
Mean	0.47	0.42	0.39	

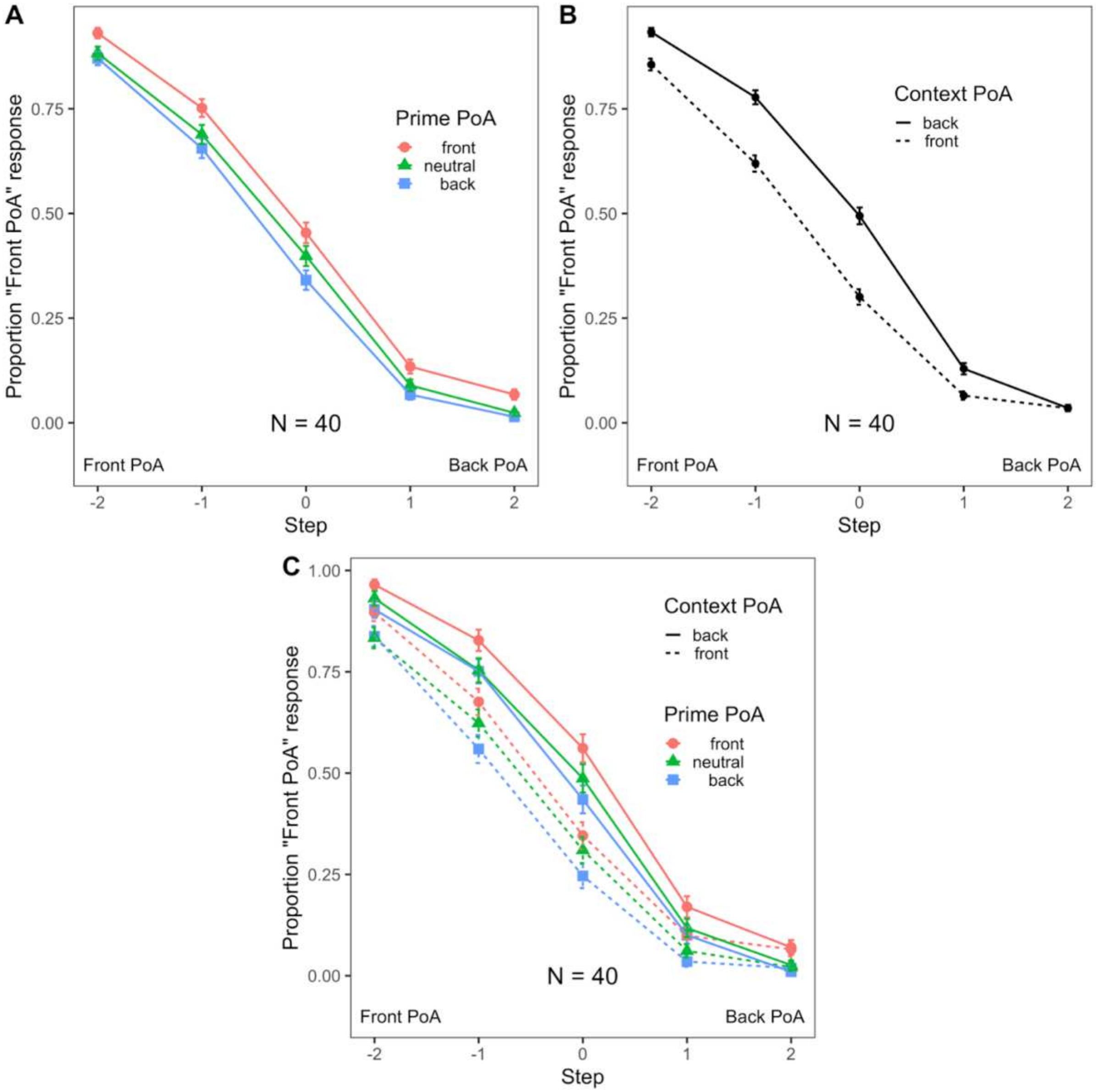

Figure 4. Responses from Experiment 2 showing decisions about the first phoneme of a target continuum (all continua were between /s/ and /ʃ/, e.g., *same*-*shame*) after a visual prime and an auditory context item. The x-axis shows continuum step, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows proportion of front PoA (/s/) responses. Colors and shapes indicate the prime that participants saw. Line type indicates context item PoA. Error bars represent 95% confidence intervals. A. Effects of prime. B. Effects of auditory context. C. Effects of both prime and context.

Our mixed effects logistic regression model revealed significant effects of Prime (χ^2 = 9.08, *p* = .011) and Context (χ^2 = 15.11, *p* < .001), indicating that participants' responses were influenced by the written prime and the PoA of the auditory context. In other words, participants made more /s/ responses after /s/-biasing primes, such as SAME, consistent with Experiment 1, and made more /s/ (front PoA) responses after context items ending in a back PoA, such as *maniac*, which is consistent with the expected CfC effect. The model also revealed a significant effect of Step (χ^2 = 23.00, *p* < .001), indicating that participants made more front-PoA responses for more front-PoA steps. None of the interactions were significant (all *p* > .05); notably, we do not find evidence that the effect of one cue (e.g., visual prime) is affected by the other cue (e.g., articulatory context).

We conducted follow-up tests to analyze pairwise comparisons for Prime, correcting for multiple comparisons. As in Experiment 1, there were more front-PoA responses for a front prime than for a neutral prime (contrast estimate: .417, *z-ratio* = 2.972, *p* = .007), and more front-PoA responses for a front prime than for a back prime (contrast estimate: .696, *z-ratio* = 3.128, *p* = .004). However, unlike in Experiment 1, the difference between front-PoA responses after a neutral prime and after a back prime was not statistically significant (contrast estimate: .280, *z-ratio* = 2.219, *p* = .056).

Discussion

In this study, we examined how two qualitatively different cues occurring at different times and in different modalities influence phoneme identification. Overall, we found that written primes and auditory contexts each impacted phoneme identification, but there was no interaction between these constraints, suggesting that their impacts were additive. Further supporting this point is a test of whether the presence of auditory context information led to different use of the prime. At the suggestion of a reviewer, we compared the effects of prime in Experiments 1 and 2. Neither the effect of experiment nor any interactions involving experiment and prime were significant in this model. Thus, at least in the case of these two specific cues, listeners' decisions about ambiguous phonemes appear to be influenced by all available constraints.

Notably, the experiment included combinations where the two constraints were in conflict (e.g., a prime biasing a listener towards "same" and an auditory context biasing a listener towards "shame") and where the prime was neutral and listeners only had auditory context. Thus, if the constraints were interactive, we might expect the size of the auditory context effect, for example, to differ based on whether or not there was priming information (neutral versus biasing primes). Alternatively, we could also find that the size of the priming effect varies based on whether there was biasing auditory information. As mentioned above, because we found no differential effects of priming across Experiments 1 and 2, this evidence also points towards additive effects. Our findings suggest that listeners use these two cues additively, such that the presence of one cue does not change the effect of the other.

The results raise important questions about how cue integration unfolds over time, which we consider more fully in the General Discussion. For instance, it is possible that the prime sets a baseline expectation that the auditory context can then shift. Temporally sensitive neural measures (such as EEG) might help us unpack exactly how this processing unfolds. Furthermore, it is noteworthy that our two cues differ in the timing of when information becomes available. Because listeners leveraged both sources of information, our results suggest that listeners can integrate multiple cues over a relatively long timespan (~1.5 sec). Exactly when information from these cues is integrated remains an open question.

Having established that listeners can combine qualitatively distinct cues, we next ask whether quantitative variation in one cue's reliability can shift the use of available cues.

Experiment 3

In Experiment 3, we examine how the reliability of visual priming influences identification of ambiguous word-word minimal pairs, with and without the presence of coarticulatory context. In Experiments 1 and 2, listeners sometimes received mismatching prime-target pairings (e.g., seeing the written word SHAME before hearing a relatively clear *same* token). Listeners are sensitive to the likelihood of cue co-occurrence and have been shown to change how they weight cues in situations that do versus do not track typical co-occurrence of the cues (Bushong & Jaeger, 2019; Idemaru & Holt, 2011). For our endpoint auditory tokens (e.g., clear *same* or *shame* tokens), clear mismatches between the visual prime and the auditory target may lead participants to consider the primes to be unreliable predictors of what they will hear. We will therefore examine how reliance on priming might change when clear conflict cases are removed (rendering primes more reliable predictors of targets). If listeners are sensitive to the obvious mismatch with inconsistent primes at endpoints, this change should enhance the effect of the written prime. If listeners do not perform differently with a more reliable prime, this may suggest that assessment of contextual reliability operates over more realistic language input (i.e., distributions of cues encountered in real-world language settings as opposed to cue distributions in our paradigms) or relates to other differences, such as the timing of disambiguating information (where visual priming and auditory context occur on different timescales).

Experiment 3 thus includes replications of Experiment 1 (Experiment 3a) and Experiment 2 (Experiment 3b) with the following change: all non-neutral primes for endpoint auditory tokens are consistent with the auditory token (e.g., participants only see SHAME or ######## before hearing the *shame* endpoint and never see SAME before hearing the *shame* endpoint, and vice versa).

Experiment 3a

Methods

Materials. Stimuli were the same as for Experiment 1.

Participants. We collected data from 63 participants in order to achieve our pre-registered target sample size of 40 participants (21 female, 18 male, 1 other/decline to state; age range: 18-34, mean age: 28). Experimental sessions took approximately 60 minutes. Participants were paid $12 for their participation, consistent with Connecticut minimum wage ($12/hr at the time). We used the same recruitment and exclusionary criteria as for Experiment 1.

Procedure. Procedures were the same as for Experiment 1 with one key change. When the auditory target item was an endpoint token (i.e., the most /s/-like or /ʃ/-like token), the written prime was never a mismatch (i.e., participants only ever saw SHAME or ######## before hearing *shame*). For each target continuum, then, endpoint tokens were presented with matching primes two-thirds of the time, and neutral primes one-third of the time (as opposed to one-third matching, one-third mismatching, and one-third neutral as in Experiment 1).

Analyses. We used the same model structure as for Experiment 1. This model included fixed effects of Prime (front, back, and neutral; sum-coded) and Step (which ranged from -2 to +2) and their interaction. Our model also included by-subject and by-target-pair random slopes for Prime and Step and their interaction, as well as by-subject and by-target-pair random intercepts, with no random correlations between slopes and intercepts. However, we restricted our analyses to only include our non-endpoint auditory targets, as the endpoints only occurred with one biasing prime and the neutral prime.

Results

Participants made more front-PoA responses when they had a front-PoA prime as compared to neutral and back-PoA primes (see Figure 5). Across the three middle steps (for which priming was balanced), participants made a front-PoA response 47% of the time they had a front-PoA prime, 41% of the time they had a neutral prime, and 38% of the time they had a back-PoA prime.

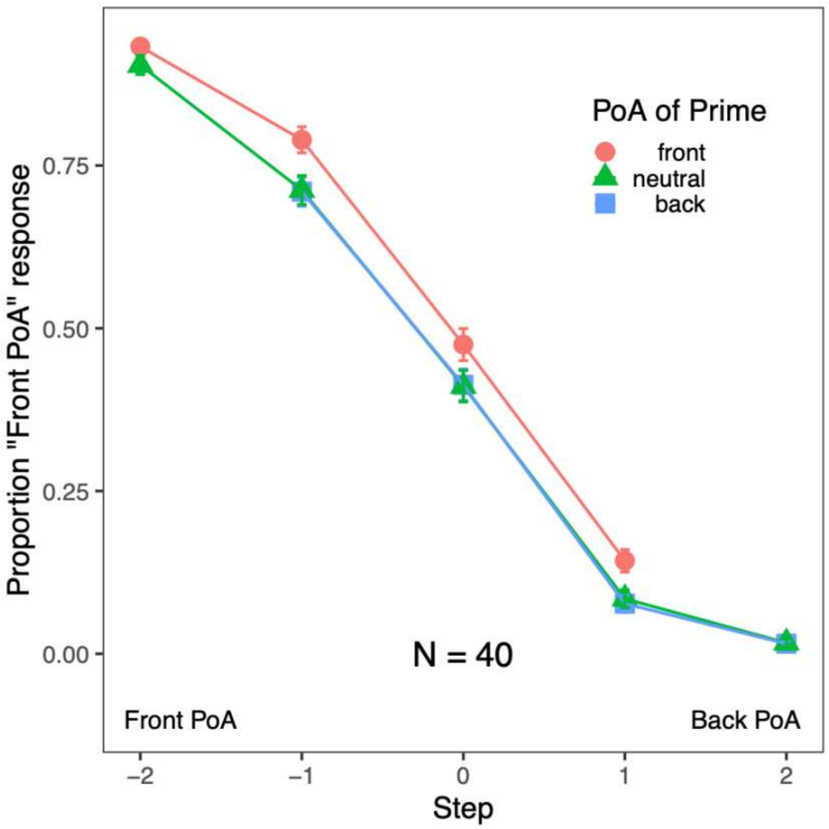

Figure 5. Responses from Experiment 3a showing decisions about the first phoneme of a target continuum (all continua were between /s/ and /ʃ/, e.g., *same-shame*) after priming, where endpoint tokens always had consistent primes. The x-axis shows the continuum steps, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows proportion of front (/s/) responses. Colors and shapes indicate the prime that participants saw. Error bars represent 95% confidence intervals. The absence of data for front primes at Step 2 and for back primes at Step -2 reflects the reliability manipulation; in contrast to Experiments 1 and 2, primes were never presented in the "high-conflict" cases where primes clearly mismatch endpoint items.

Our mixed effects logistic regression model revealed a significant effect of Prime ($\chi^2 = 15.20$, $p < .001$), indicating that participants' responses were influenced by the written prime.

The model also revealed a significant effect of Step (χ^2 = 18.80, *p* < .001), indicating that participants made more front-PoA responses for more front-PoA steps. The interaction between Prime and Step was not significant (χ^2 = 0.82, *p* = .66), suggesting that there was no difference in the effect of priming across continuum steps.

We conducted follow-up tests to analyze pairwise comparisons for Prime. There were more front-PoA responses for a front prime than for a neutral prime (contrast estimate: .486, *z-ratio* = 4.986, *p* < .001), and more front-PoA responses for a front prime than for a back prime (contrast estimate: .516, *z-ratio* = 3.170, *p* = .003). However, there were no differences in responses after a back prime and after a neutral prime (contrast estimate: .030, *z-ratio* = 0.312, *p* = .936).

Examining the Effect of Reliability with One Cue. To examine how reliability affected the use of priming, we compared results from Experiment 1 to Experiment 3a (restricting analyses to exclude endpoint tokens). These analyses were not pre-registered and hence are exploratory. Our model followed a similar structure as the main model for Experiment 3a, with the addition of a fixed effect of Reliability (Experiment 1: unreliable, Experiment 3a: reliable; sum-coded) and its interactions with Prime and Step, and additional by-target-pair random slopes for Reliability and its interactions with Prime and Step. Models for reaction time data can be found in supplementary materials.

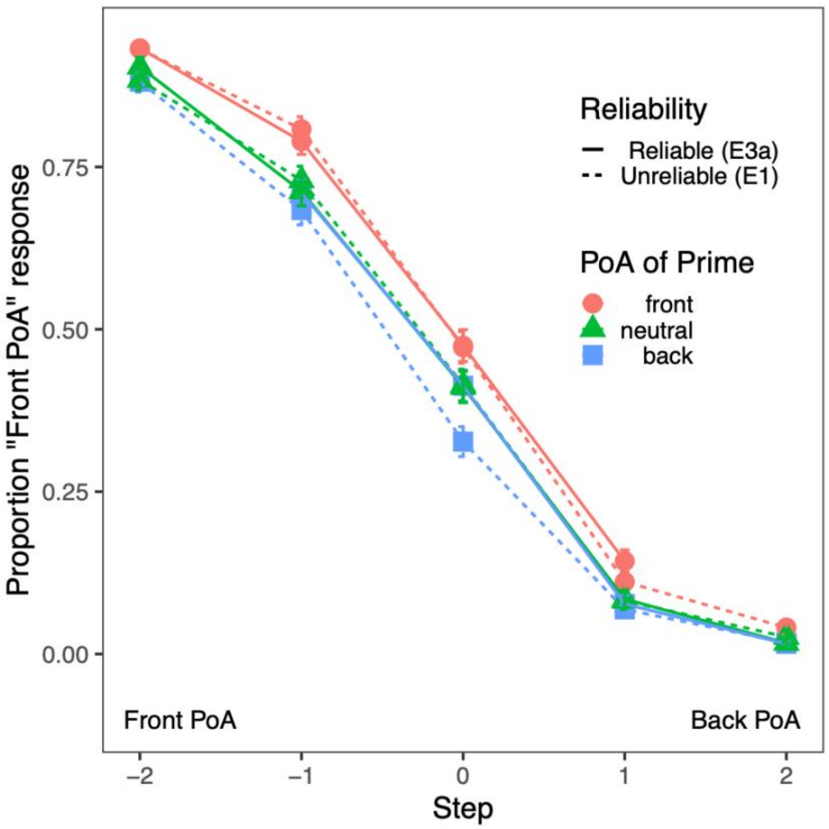

Figure 6. Response data for Experiment 1 (unreliable prime-target relationship) and Experiment 3a (reliable prime-target relationship). We were interested in whether responses would differ based on the reliability of the prime and target relationship. The x-axis shows continuum step, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows proportion of front PoA (/s/) responses. Colors and shapes indicate the prime that participants saw, and line type indicates the reliability condition. Error bars represent 95% confidence intervals.

Response Data. Because response data for Experiment 1 and Experiment 3a has been analyzed and presented in results above, we will focus only on effects in the model that relate to Reliability (for a plot of both datasets together see Figure 6). Our mixed effects model revealed a significant Prime by Reliability interaction (χ^2 = 9.52, p = .009), which appears to be attributable to the lack of a back-neutral difference in the Reliable condition (Experiment 3a). No other effects or interactions involving Reliability were significant (all p > .05).

Discussion

We do not find evidence that increasing the reliability of the priming information (i.e., only presenting consistent primes before clear endpoint tokens) increases reliance on priming information when hearing ambiguous auditory tokens. While we still observe an overall effect of prime (and importantly a difference in responses between front and back primes) for the ambiguous tokens, participants were no more likely to respond in accordance with the prime when endpoint tokens were only paired with consistent primes (Experiment 3a) than when endpoint tokens were also paired with mismatching primes (Experiment 1). It is possible that this reflects a ceiling effect; the priming effect may have already been maximally large, and so increasing reliability did not make any difference. It is also possible that we did not change endpoint-priming consistency sufficiently to make listeners reweight reliance on the prime (given that we only manipulate the reliability of endpoint steps, whereas Bushong & Jaeger [2019] manipulated reliability across all steps). Future work with a more graded priming manipulation (e.g., where Steps 1 and -1 are not presented equally with both primes) could test this possibility. Future work could also explore whether changing the reliability of priming in the opposite direction (i.e., making primes less reliable or even completely unreliable) would affect the impact of primes.

Experiment 3b

The goal of Experiment 3b was to replicate Experiment 2, with the addition of the same reliability manipulation used in Experiment 3a. Namely, all non-neutral primes for endpoint auditory tokens matched the token.

Methods

Materials. Stimuli were the same as for Experiment 2.

Participants. We collected data from 55 participants in order to achieve our pre-registered target sample size of 40 participants (25 female, 15 male; age range 18-34, mean age: 27). Experimental sessions took approximately 60 minutes. Participants were paid $12 for their participation, consistent with Connecticut minimum wage ($12/hr at the time). We used the same recruitment and exclusionary criteria as in Experiment 1.

Procedure. Procedures were the same as for Experiment 2 with the same key change as in Experiment 3a: when the auditory target item was an endpoint token (i.e., the most /s/-like or /ʃ/-like token), the written prime was never a mismatch (i.e., participants only saw SHAME or ######## before hearing *shame*). For each target continuum, then, endpoint tokens were presented with matching primes two-thirds of the time, and neutral primes one-third of the time (as opposed to one-third matching, one-third mismatching, and one-third neutral as in Experiment 2).

Analyses. Analyses followed the same structure as for Experiment 2.

Results

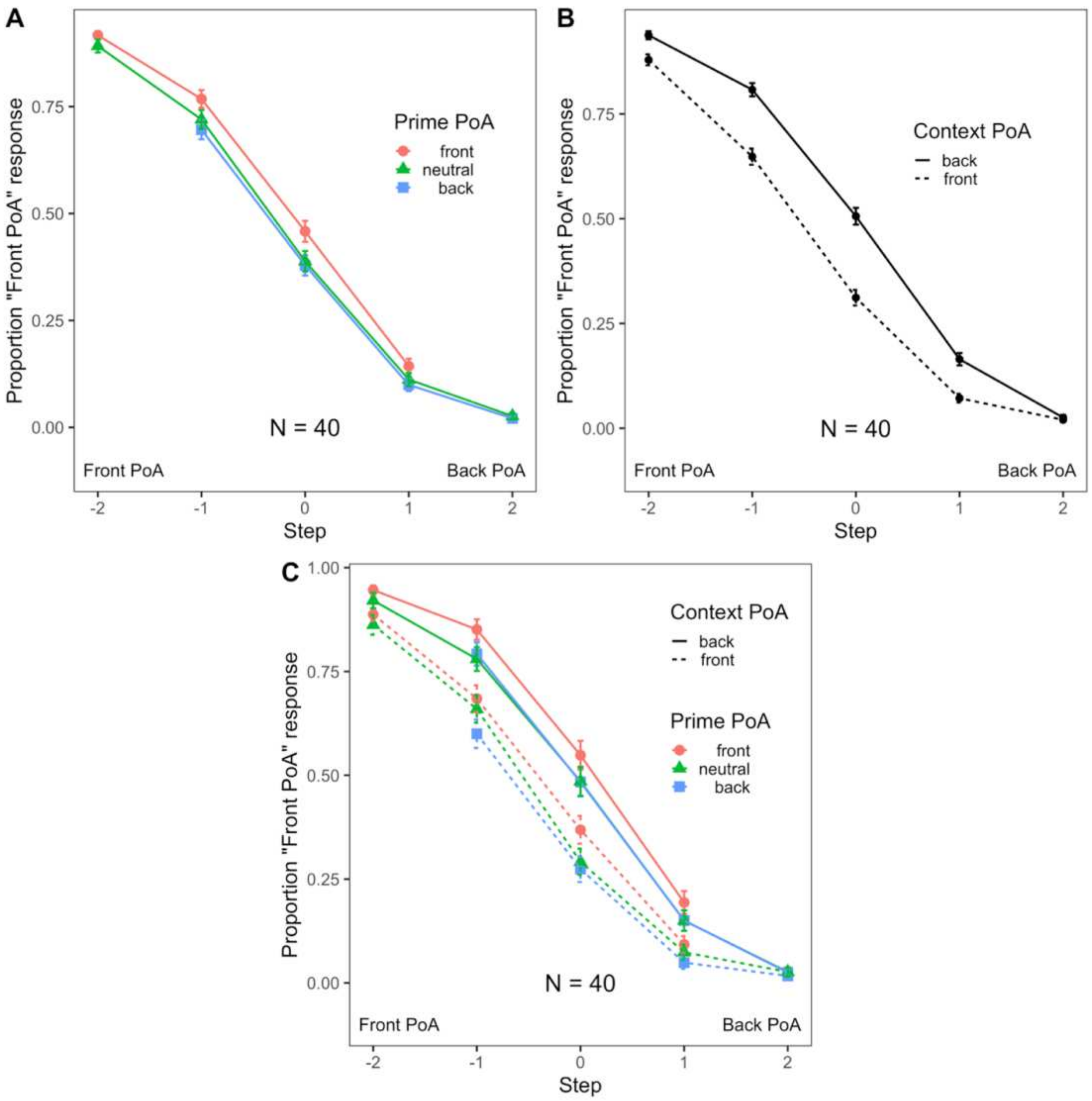

Figure 7. Responses from Experiment 3b showing decisions about the first phoneme of a target continuum (all continua were between /s/ and /ʃ/, e.g., *same*-*shame*) after a visual prime and an auditory context item, where endpoint tokens always had consistent visual primes. The x-axis shows continuum step, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows proportion of front (/s/) responses. Colors and shapes indicate the prime that participants saw. Line type indicates context item PoA. Error bars represent 95% confidence intervals. A. Effect of prime. B. Effect of auditory context. C. Effects of both prime and context.

As shown in Figure 7, participants were influenced by both the prime and the auditory context items. Across the three middle steps, participants made more front-PoA responses when they heard an auditory context item that ended in a back-PoA (overall 49%) than when the context item ended in a front-PoA (overall 34%). Likewise, they made more front-PoA responses when they saw a front prime (overall 46%) than when they saw a neutral prime (overall 41%) or a back prime (overall 39%). Breakdowns of the proportion of front responses by condition can be found in Table 2.

Table 2.

Proportion of front-PoA responses based on Context and Prime.

Context	Prime			
	Front	Neutral	Back	Mean
Back	0.53	0.47	0.48	0.49
Front	0.38	0.34	0.31	0.34
Mean	0.46	0.41	0.39	

Our mixed effects logistic regression model revealed significant effects of Prime (χ^2 = 15.37, *p* < .001) and Context (χ^2 = 17.56, *p* < .001), indicating that participants' responses were influenced by the written prime and the place of articulation of the auditory context. The model also revealed a significant effect of Step (χ^2 = 19.92, *p* < .001), indicating that participants made more front-PoA responses for more front-PoA steps. None of the interactions were significant (all p > .05).

We conducted follow-up tests to analyze pairwise comparisons for Prime. There were more front-PoA responses for a front prime than for a neutral prime (contrast estimate: .387, *z-ratio* =

4.742, *p* < .001), and more front-PoA responses for a front prime than for a back prime (contrast estimate: .516, *z-ratio* = 4.348, *p* < .001). However, the difference between front-PoA responses after a neutral prime and after a back prime was not statistically significant (contrast estimate: .128, *z-ratio* = 1.610, *p* = .225).

Examining the Effect of Reliability with Multiple Cues. To examine how reliability affected the impact of priming, we compared results from Experiment 2 to Experiment 3b (restricting analyses to exclude endpoint tokens). These analyses were not pre-registered and hence should be interpreted as exploratory. Our model followed a similar structure as the main model for Experiment 3b, except that we did not include Context as a factor (as we were interested in the effect of prime and were likely not powered for a four-way interaction) and instead included a fixed effect of Reliability (Experiment 2: unreliable, Experiment 3b: reliable; sum-coded) and its interactions with Prime and Step, and additional by-target-pair random slopes for Reliability and its interactions with Prime and Step. Models for reaction time data can be found in supplementary materials.

Response Data. Because response data for Experiment 2 and Experiment 3b has been analyzed and presented in results above, we will focus only on effects in the model that relate to Reliability (for a plot of both datasets together see Figure 8). No effects or interactions involving Reliability were significant (all p > .05).

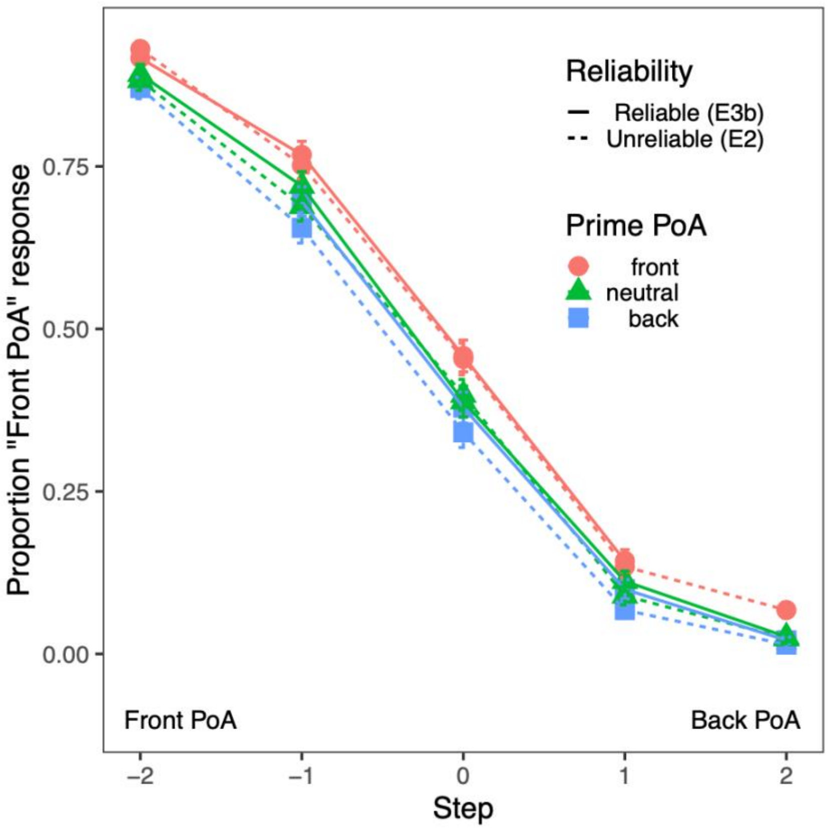

Figure 8. Response data for Experiment 2 (unreliable) and Experiment 3b (reliable), based on Prime (collapsed across auditory Context). The x-axis shows the continuum steps, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows proportion of front PoA (/s/) responses. Colors and shapes indicate the prime, and line type indicates reliability condition. Error bars represent 95% confidence intervals.

Discussion

Similar to the results presented in Experiment 3a, we find that increasing the reliability of the priming information (i.e., only presenting consistent primes before clear endpoint tokens) did not increase the effect of prime at the level of the response data. Even though we did not find a reliability effect in Experiment 3a (with just one cue), it was possible in Experiment 3b that we would observe an effect of reliability in the presence of two cues, but we did not. It is possible that making prime an even more reliable cue might shift the overall weighting of cues, such that we would observe more reliance on the prime (and subsequently less reliance on the auditory context). Alternatively, as we discussed with respect to Experiment 3a, the lack of reweighting observed in Experiment 3b is consistent with the possibility that listeners were already maximally impacted by the prime.

As an additional test of whether we might see a different effect of the prime depending on whether or not there was auditory context information, we compared the effects of prime in Experiments 3a and 3b (a parallel analysis to that which was discussed in the Discussion of Experiment 2). Neither the effect of experiment nor any interactions involving experiment and prime were significant in this model. Thus, at least in the case of these two specific cues, listeners' decisions about ambiguous phonemes appear to be influenced by all available constraints.

General Discussion

Overall, the goal of this project was twofold: to investigate how people make use of (or are impacted by) qualitatively different cues when identifying phonemes and to assess how the reliability of a cue influences its use (or impact). As we discussed in the Introduction, most work on constraint integration has focused on one constraint at a time, or qualitatively similar constraints (e.g., pairs of acoustic-phonetic cues in trading relations studies; Repp, 1982). We know less about whether or how listeners integrate qualitatively distinct kinds of constraints or reconcile them when they conflict or vary in reliability. In the following subsections, we will briefly discuss our three key findings: (1) that resolution of acoustic-phonetic ambiguities can be influenced by visual primes that precede the point of ambiguity by more than a second; (2) that when we crossed two qualitatively distinct constraints (visual priming and coarticulatory context), we observed additive effects (rather than interactions, or differential use of one constraint in the presence of the other); and (3) that this additivity persisted even when we enhanced the reliability of the visual prime (by never presenting primes that would conflict with unambiguous endpoint tokens). After discussing each of these major findings, we will turn to the implications for theories of language processing.

Identity Priming Influences Phoneme Identification

We first established that visual identity priming influences identification of ambiguous phonemes in Experiment 1. While this finding was expected, we note that to our knowledge this study provides the first such test of the influence of visual priming on ambiguous phonemes. Previous work has found that (semantic) primes exert an early influence on acoustic-phonetic processing. Getz and Toscano (2019) used a paradigm with a visual prime that was semantically related to one end of an ambiguous voice onset time (VOT) minimal pair (e.g., with a token ambiguous between *park* and *bark*, AMUSEMENT would be a semantic prime for *park*). Getz and Toscano (2019) found that semantic primes do influence perception of ambiguous tokens, and that this (semantic) influence occurs very early: about 100ms after the onset of the target stimulus, listeners are already more likely to perceive the onset phoneme as consistent with the prime (as indexed by the N100 ERP component, which has been shown to reflect perceptual information such as VOT linearly; Toscano et al., 2010). For example, if AMUSEMENT precedes the auditory token *[?]ark*, the N100 for the ambiguous token more closely resembles the N100 for a clear /p/ token than if KITCHEN precedes *[?]ark*.

Since priming can serve as a proxy for prediction, by creating a possibly implicit expectation (Blank & Davis, 2016; Sohoglu et al., 2014), our finding could serve as a basis for examining how prediction strength influences perception. Because Getz and Toscano (2019) demonstrated that semantic priming affects perception of ambiguous phonemes in behavioral and ERP measures while we showed that visual priming affects perception of ambiguous phonemes in behavioral measures, future work could investigate how visual priming affects ERP measures. To what degree, for example, would the shift in N100 that results from a semantic prime (Getz & Toscano, 2019) reflect the strength of the prediction? By varying levels of relatedness between the prime and the ambiguous target (from unrelated, to various degrees of semantic association, to identity priming, where the prime and target are identical), future work could examine how the spectrum of relatedness between prime and target is reflected in the N100.

Other manipulations including a more comprehensive reliability manipulation (Bushong & Jaeger, 2019) or introducing visual noise (to manipulate the perceptual reliability of primes) could allow for further exploration of how the degree of prediction from prior knowledge affects perception. Such studies would allow us to understand more about the nature of early perceptual encoding, assessing whether the shift in the N100 reflects the degree of perceptual change the participant experiences.

Use of Qualitatively Distinct Constraints

When listeners had both a visual prime and acoustic coarticulatory context available to constrain resolution of an ambiguous fricative, we observed an additive effect of both cues. Because these cues were qualitatively distinct (occurring in different modalities with different temporal relations to the point of ambiguity), it was quite possible that we might observe differential use of one source of information in the presence of the other. The auditory context immediately precedes the target token and exists in the same modality. On the other hand, the written prime was explicit and occurred first, which could have made it the more salient cue (particularly in Experiment 3, where clear mismatches were removed).

Reaction time analyses (see Supplementary Materials) also support the idea that listeners are independently sensitive to the different types of information. Namely, pre-registered reaction time analyses indicated that participants were faster when the target was preceded by any visual word prime (e.g., faster responses to the target following a prime of SIGN or SHINE relative to a neutral prime ########) as well as faster for unambiguous continuum steps (a clear auditory *sign* or s*hine*) compared to an ambiguous auditory stimulus. These findings support the idea that having any priming information made participants faster, whether or not this information was in conflict with the auditory coarticulatory information, suggesting independent processing across priming and coarticulatory sources of information.

The finding of additivity of constraints is also in line with findings from Lai et al. (2022), who similarly found additive effects of lexical and coarticulatory constraints. However, the present work differs from the study by Lai et al. (2022) in several key ways. Though the nature of the coarticulatory constraints was similar in the two studies, lexical constraints differed in modality (auditory in Lai et al.'s study, visual in the current work) and in the timing of lexical information (in Lai et al's study, lexical status could only be determined after the lexical uniqueness point, whereas in the current work, visual primes were provided ~1 sec before the auditory target). It is striking that both studies observed additive effects of lexical information and coarticulatory context, despite differences in the modality and timing of lexical information. It is particularly noteworthy that we still observed additive effects of these cues in Experiment 3, when we changed the reliability of the cues. Whether these results will hold over different combinations and different numbers of cues remains an open question. Nonetheless, the current findings might be extended to questions involving individual differences that affect the impact or availability of constraints as well as questions about the mechanisms of cue integration across different types of cues. We consider both of these possible extensions.

Individual Differences

The results of Lai et al. (2022) provide some evidence for a trade-off between reliance on high-level contextual information and reliance on low-level acoustic information. In their study, an analysis of individual differences indicated that listeners who relied strongly on lexical knowledge relied relatively less on coarticulatory cues, and vice versa. Non-preregistered analyses of our data (see Supplemental Materials) found a similar pattern in Experiment 2 (i.e., that listeners who relied strongly on lexical information relied relatively less on coarticulatory information, and vice versa). Notably, however, there was no such relationship in Experiment 3 (where we manipulated the reliability of the cues). However, the analysis for Experiment 3 only considered non-endpoint continuum steps, and strikingly, neither the correlation in Lai et al. (2022) nor the correlation in our Experiment 2 are significant when only middle steps are included. Thus, the two studies provide some evidence for a tradeoff in how strongly listeners rely on lexical knowledge and coarticulatory information, but that tradeoff might be driven primarily by continuum endpoints where one cue (acoustics) is likely to dominate.

There are many ways in which these effects might vary across different types of listeners. Older adults, for example, have been shown to rely more on higher-level contextual information (e.g., lexical information as opposed to acoustic information; Mattys & Scharenborg, 2014). Interestingly, however, work from Luthra, Peraza-Santiago, Saltzman, and colleagues (2021) has demonstrated that older adults do not show a larger influence of lexical information to resolve phonetic ambiguities in the case of coarticulation. In other words, it does not seem to be the case that those who potentially rely more on lexical information show larger effects of CfC, suggesting that the CfC effect is perhaps robust but already at ceiling in the average listener (such that those who in general rely more on higher level contextual constraints do not show larger CfC effects). The fact that we observe additive effects of auditory context and written primes, however, suggests that larger perceptual shifts are still possible with additional sources of information. Future work could examine whether older adults (or others who tend to rely more contextual information; see Crinnion et al., 2021 and Kaufeld et al., 2020) demonstrate larger effects of the prime (at least when biasing information is consistent) than younger adults (or individuals who rely more on the acoustic signal).

Previous research has also suggested that variation in the impact of acoustic information (e.g., sensitivity to subphonemic information; Li et al., 2019) and the impact of lexical information (e.g., relative reliance on lexical vs. acoustic information; Giovannone & Theodore, 2021) is related to variation in language abilities (e.g., phonological skills, expressive and receptive language abilities). We would predict, then, that individual differences in language ability might influence relative weighting of cues. In a situation where individuals have two sources of information that could potentially influence interpretation of acoustic cues, it would be interesting to test whether those with weaker language skills rely more on the prime (as opposed to the auditory context, which is an acoustic cue), since those with weaker language abilities tend to rely more on lexical-level information as opposed to acoustic information (Giovannone & Theodore, 2021; though see Li et al., 2019 for evidence of individuals with lower language abilities relying more on acoustics).

Finally, it is important to consider how listeners' tendencies to use acoustic cues may influence how they make use of multiple sources of information. Work from Kapnoula and colleagues (2017) demonstrated that more gradient listeners (i.e., listeners who, when asked how /s/- or /ʃ/-like a certain token sounds, respond in a less binary fashion) are more likely to use a secondary acoustic cue. In the current paradigm, we might expect more gradient listeners to be influenced more by the auditory context than the written prime. It also could be possible, however, that they would be overall more likely to use *any* type of additional information.

Mechanisms of Cue Integration

Future work can also examine the mechanisms behind additive cue use in speech perception. First, understanding this phenomenon more broadly requires that other combinations of cues be tested. While the fact that our cues were quite distinct from each other might suggest that any pair of cues would result in additive effects, this remains an open question. Second, it is important to consider that the current work cannot speak to *how* listeners use both cues in this experiment. Particularly because the timing between our two cues varies considerably (e.g., the visual prime appears early, over 1s before the target, whereas the auditory context immediately precedes the target), examining the time course of processing will be important for future work.

Using eye tracking or EEG would make this possible and provide a clearer picture of how and when the observed additivity arises. In fact, eye tracking work looking at integration of acoustic-phonetic cues and even lexical and phonetic cues suggests that information influences processing as soon as it is available (Kingston et al., 2016; McMurray et al., 2008; Reinsch & Sjerps, 2013; Toscano & McMurray, 2015; see also Dahan et al., 2001, and Li et al., 2019). To create a visual world analog of our paradigm, we could present participants with visual referents for the target words (e.g., a sun for *shine* and a stop sign for *sign*; see Kaufeld et al., 2020 for a similar approach), and examine where people looked as the different cues unfolded (perhaps an auditory identity prime to prevent visual interference with visual referents and then the auditory context) which may shed light on whether, over the course of a given trial, listeners tend to rely on just one source of information during that trial (i.e., an early commitment) or if they continue to consider both sources of information (once both sources of information are available).

ERP analysis of the N100 could also shed light onto how predictions from both the prime and the auditory context are integrated. By comparing ERPs on trials with just one source of information to trials with two (potentially competing) sources of information we can see whether the N100 response is similar whether there are two sources of constraint vs. just one, or if it demonstrates something analogous to summation (for similar approaches see Getz & Toscano, 2019 and Noe & Fischer Baum, 2020). Using ERPs could also allow us to examine whether the locus of impact for both cues is perceptual (i.e., evident at the N100) or if one has a later (potentially post-perceptual/decision stage) impact.

Cue Reliability

In Experiment 3, we increased the reliability of the written prime to test whether listeners would use the prime more than they did in Experiments 1 and 2. Findings from Bushong and Jaeger (2019) suggest that when acoustic information tracks more realistically with lexical information in the sentence context (i.e., when unambiguous endpoint tokens, such as *dent,* are always paired with related sentence contexts, such as *dent in the fender*, but never with unrelated contexts, such as *dent in the forest*), listeners rely more on that lexical information. Hence, in our studies we expected to find that when visual lexical priming was more consistent with the acoustic information (by never pairing visual primes with the clear endpoints they conflict with, in Experiments 3a and 3b), listeners would rely more on the prime. Furthermore, work from Idemaru and Holt (2011) suggests that when two cues track in a more naturalistic way, listeners are more likely to use the secondary cue (F0) when the cue they would typically use (VOT) is ambiguous than if the correlation of the two cues was opposite that of natural speech. These findings again suggest that we might expect listeners to use the prime more when acoustic information tracked more consistently with it, as in Experiment 3a and even in Experiment 3b (when listeners had yet another cue to use). However, differences in cue weighting between more reliable and less reliable cases typically reveal downweighting in the unreliable cases, but often find no difference between neutral and reliable relationships between two cues (e.g., Idemaru & Holt, 2011). Furthermore, Kim et al. (2020) found that in order to boost reliance on a weaker cue, the primary cue needed to be made less reliable.

We did not see such an effect of our reliability manipulation. Because our experiment used a small set of repeated stimuli, it is possible that listeners were already maximally impacted by visual priming. Increasing variation in the stimuli or adding noise to the stimuli might result in a detectable impact of the reliability manipulation. It is also possible that we might see changes in priming impact with a more graded priming manipulation (where not only the endpoints but all steps were paired with their more likely prime in proportion to their ambiguity). Our task (two alternative forced choice) may have limited the types of effects we were able to detect, such as any information about variation in how /s/-like participants perceived a given token. It could also be interesting to examine whether we might observe effects of reliability with a different behavioral task, namely, one that asked participants to rate how /s/- or /ʃ/-like a certain token sounds. Additionally, future work should explore whether changes in reliability (perhaps introduced by creating *less* reliable priming information, for example) differentially affect scenarios in which there is just one cue to rely on (as in Experiments 1 and 3a) as opposed to multiple sources of information (as in Experiments 2 and 3b). This line of work would perhaps mirror that of work that shows that reliance on a secondary cue changes when it is made less reliable, as in Idemaru and Holt (2011).

Implications

Our current findings add to the growing body of literature suggesting that (at least at the group level), listeners demonstrate patterns of additivity across information provided from different cues. As discussed above, building a more comprehensive mechanistic account of cue integration across different types of cues is a necessary step for the field.

Many frameworks for language processing, such as interactive activation models (e.g., TRACE; McClelland & Elman, 1986), autonomous models (e.g., Shortlist B; Norris & McQueen, 2008), and other Bayesian frameworks (e.g., Ideal Adapter; Kleinschmidt & Jaeger, 2015), have mechanisms for implementing effects of information from different types of cues. TRACE can simulate phonetic context effects, and the original and current implementations (jTRACE; Strauss et al., 2007) include (mainly unexplored) facilities for simulating priming. In a Bayesian model, different prior distributions can be explored (for example).

However, the paradigms we used here are not easy to simulate with current models of spoken word recognition. In TRACE for example, priming is implemented by the researcher pre-setting the resting level activation value for a given word, but it is not clear what impact a spoken word intervening between the prime and target should have. Similarly, in a Bayesian framework, one could assume a system retains information from a prime until it becomes relevant and then integrates it at the appropriate point in the stimulus. A problem with either approach is that they could become exercises in data fitting, rather than providing a general processing framework (since the researcher simply stipulates that priming happens and fits a parameter for the magnitude of the prime, but neither approach provides an integrated way to simulate the actual phenomenon of priming, and especially impact from temporally distant primes; priming appears to be outside the explanatory scope of current models).[1]

In the current work, for example, the prime occurred over one second before the ambiguous auditory information and yet was reliably used to resolve this ambiguity. A better understanding of the temporal proximity of various sources of information is needed for a more complete model of speech perception. It could be the case, for instance, that the high degree of ambiguity in the current experiments (i.e., listeners heard many ambiguous tokens, perhaps more than are typically encountered in naturalistic listening environments) broadened the window of time over which listeners were willing to consider information as relevant for resolving the ambiguities in the speech. On our view, future research seeking to build a more mechanistic understanding of the types of information that may be considered should directly consider the *functional* time window of integration, which may be context dependent. It also could be that under higher degrees of uncertainty, more temporally distant cues are considered. Varying the amount of uncertainty and manipulating the time window of previous information (and potentially intervening information) would allow us to assess whether the system broadens its window of influence if there is more uncertainty present. The possibility that the temporal window of cue integration expands under conditions of ambiguity is an intriguing one, especially

[1] There are models of sentence processing where the processing of one word can influence the processing of another (e.g., via continuous updating of semantic representations; Rabovsky et al., 2018). While these models do not simulate processing at the level of speech – with continuous input and acoustic-phonetic constraints – they might provide inspiration for extending current models of human speech processing.

since it would pose important constraints on the underlying architecture of the speech processing system as well as cognitive theories of speech perception.

Another important issue for integration processing frameworks to consider is that of the specific type of information involved. There are, for example, instances of modeling approaches that explain how different acoustic cues are weighted in the perception of speech sounds (Crinnion et al., 2020; McMurray & Jongman, 2011) and even accounts of how a system might compute these cues relative to other types of contexts (e.g., who is talking; McMurray & Jongman, 2011). However, understanding from a neurocomputational perspective, what function different types of information serve is important (Luthra, Li, et al., 2021). For example, do higher level sources of information support prediction, upon which changes in neural activity would reflect some error signal (or deviation from that prediction), or do they serve as sources of information that boost activation of consistent signals?

A first step towards answering some of these questions will come from more extensive exploration of cue additivity across a variety of constraints. As suggested by a reviewer, a *metastudy* approach (a meta-analysis approach leveraging Bayesian hierarchical modeling; Baribault et al., 2018; DeKay et al., 2022) could be particularly powerful in this domain. Pooling data from many studies could allow for fruitful exploration of how qualitatively different cues (from different levels of linguistic representation) and quantitatively different cues (due to cue strength, reliability, or timing, for example) are integrated in speech processing.

Conclusion

In everyday speech, listeners often have many sources of information available (e.g., who is talking, what the general conversational topic is, how fast the person is talking, etc.) that can potentially help resolve acoustic ambiguities. While much research has examined how different types of information can, *individually*, influence identification (and arguably, perception) of ambiguous speech sounds, how listeners make use of *multiple* (sometimes conflicting) sources of information for resolution of a single ambiguity remains underspecified.

A challenge for existing computational frameworks will be to simulate the additive effects we observed despite differences in modality and timing and to likewise consider the predictions that would then arise – for instance, considering shifts in cue reliability. Finally, to what extent are listeners actually influenced by all available information, and when might the impact of different cues change? Moreover, how consistent are such patterns across individuals? While there are many remaining questions, the findings reported here provide a first step in considering how two cues that are qualitatively distinct (in timing and modality) can simultaneously and independently influence language processing.

Declarations

Open Practices Statement

All experiments were preregistered at https://osf.io/6kmub and the data and code for all experiments are available at: https://osf.io/hzk49/?view_only=04dadaf24eb047329af4a22338a17a1a

Funding

This research was supported by National Science Foundation (NSF) grants NRT 1747486 and PAC 2043903 (PI JSM). Phoebe Gaston was supported by NSF grant CRCNS 2207770. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Anne Marie Crinnion and Phoebe Gaston were supported by NIH T32 DC017703. Sahil Luthra was supported by NIH F32 DC020625. Anne Marie Crinnion was supported by F31 DC021372. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported in part by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through project PID2020-119131GB-I00 (BLIS).

Conflicts of interest

The authors have no conflicts of interest to report.

Ethics approval

All procedures were approved by University of Connecticut's Institutional Review Board (IRB) and all participants provided informed consent.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *Journal of Experimental Psychology: General, 128*(1), 32.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, 115*(11), 2607–2612.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 5*(2), 252.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology*, *14*(11), e1002577.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat:" The effects of semantic and acoustic information on phoneme categorization. *The Journal of the Acoustical Society of America, 103*(5), 2670-2676.

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience, 39*(38), 7564-7575.

Bushong, W., & Jaeger, T. F. (2019). Dynamic re-weighting of acoustic and contextual cues in spoken word recognition. *The Journal of the Acoustical Society of America*, *146*(2), EL135–EL140.

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1084.

Crinnion, A.M., Heffner, C. C., Myers, E.B. (2021, November). Individual differences in signal-based vs. knowledge-based cues for speech perception. Poster presented at the *62nd Annual Meeting of the Psychonomic Society*, Virtual Conference.

Crinnion, A. M., Malmskog, B., & Toscano, J. C. (2020). A graph-theoretic approach to identifying acoustic cues for speech sound categorization. *Psychonomic Bulletin & Review*, *27*, 1104-1125.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975-979.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes, 16(5-6), 507-534.

DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating Psychological Science With Metastudies: A Demonstration Using the Risky-Choice Framing Effect. *Perspectives on Psychological Science, 17*(6), 1704–1736.

Denes, P. (1955). Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America*, *27*(4), 761-764.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.*, *55*, 149–179.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*(2), 143-165.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*(4), 481-496.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological Evidence for Top-Down Lexical Influences on Early Speech Perception. *Psychological Science*, *30*(6), 830–841.

Giovannone, N., & Theodore, R. M. (2021). Individual Differences in the Use of Acoustic-Phonetic Versus Lexical Cues for Speech Perception. *Frontiers in Communication*, *6*, 120.

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science*, *17*(1), 42–46.

Holcomb, P. J., & Grainger, J. (2006). On the time course of visual word recognition: An event-related potential investigation using masked repetition priming. *Journal of Cognitive Neuroscience, 18*(10), 1631-1643.

Hutchinson, K. M. (1989). Influence of sentence context on speech perception in young and older adults. *Journal of Gerontology, 44*(2), P36-P44.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1939-1956.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594.

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (20190725). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(3), 549.

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. Journal of Phonetics, 81, 100984.

Kingston, J., Levy, J., Rysling, A., & Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. Journal of Experimental Psychology: Human Perception and Performance, 42(12), 1969–1988.

Kinoshita, S. (2006). Additive and interactive effects of word frequency and masked repetition in the lexical decision task. *Psychonomic Bulletin & Review*, *13*(4), 668–673.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help?. *Language, Cognition and Neuroscience, 34*(1), 43-68.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of*

Memory and Language, 56(1), 1-15.

Lai, W., Wade, L., & Tamminga, M. (2022). Individual differences in simultaneous perceptual compensation for coarticulatory and lexical cues. *Linguistics Vanguard*, *8*(1), 29-39.

Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience, 35*(18), 7203-7214.

Lenth, R. V. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.2.

Li, M. Y., Braze, D., Kukona, A., Johns, C. L., Tabor, W., Van Dyke, J. A., ... & Magnuson, J. S. (2019). Individual differences in subphonemic sensitivity and phonological skills. *Journal of Memory and Language, 107*, 195-215.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*(3), 172-187.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

Luthra, S., Li, M. Y. C., You, H., Brodbeck, C., & Magnuson, J. S. (2021). Does signal reduction imply predictive coding in models of spoken word recognition? *Psychonomic Bulletin & Review*, *28*(4), 1381–1389

Luthra, S., Peraza-Santiago, G., Beeson, K., Saltzman, D., Crinnion, A. M., & Magnuson, J. S. (2021). Robust Lexically Mediated Compensation for Coarticulation: Christmash Time Is Here Again. *Cognitive Science*, *45*(4), e12962.

Luthra, S., Peraza-Santiago, G., Saltzman, D., Crinnion, A. M., & Magnuson, J. (2021).

Lexically-Mediated Compensation for Coarticulation in Older Adults. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *The Journal of the Acoustical Society of America*, *69*(2), 548–558.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Mattys, S. L., & Scharenborg, O. (2014). Phoneme categorization and discrimination in younger and older adults: A comparative analysis of perceptual, lexical, and attentional factors. *Psychology and Aging*, *29*(1), 150.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1064–1071.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219.

McRae, K., & Matsuki, K. (2013). Constraint-based models of sentence processing. *Sentence*

Processing, *519*, 51-77.

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition, 39*(6), 1085-1093.

Noe, C., & Fischer-Baum, S. (2020). Early lexical influences on sublexical processing in speech perception: Evidence from electrophysiology. *Cognition*, *197*, 104162.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355-376.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nature Human Behaviour, 2(9), 693–705.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92*(1), 81.

Repp, B. H., & Mann, V. A. (1981). Perceptual assessment of fricative–stop coarticulation. *The Journal of the Acoustical Society of America*, *69*(4), 1154–1163.

Repp, B. H., & Mann, V. A. (1982). Fricative–stop coarticulation: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *71*(6), 1562–1567.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. Journal of Phonetics, 41(2), 101–116.

Schneider, B. A., Li, L., & Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology, 18*(07), 559-572.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2015). afex: Analysis of factorial experiments. *R Package Version 0.13–145*.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass, 3*(1), 236-264.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2014). Top-down influences of written text on perceived clarity of degraded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 186.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, *39*(1), 19-30.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 431.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience*, *30*(5), 529-543.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous Perception and Graded Categorization: Electrophysiological Evidence for a Linear Relationship Between the Acoustic Signal and Perceptual Encoding of Speech. *Psychological Science*, *21*(10), 1532–1540.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. In C. Clifton, Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on Sentence Processing* (pp. 155–179). Lawrence Erlbaum Associates, Inc.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1005.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The Learning Signal in Perceptual Tuning of Speech: Bottom Up Versus Top-Down Information. *Cognitive Science*, *45*(3), e12947.

Supplementary Materials

Experiment 1 Reaction Time Analysis

For our analysis of reaction time data, we implemented a mixed effects model with the same fixed and random effects described for response data and linked our data to a Gamma distribution with an identity function as specified by Lo and Andrews (2015). This allows for an analysis of untransformed response time data while specifying an appropriate (non-Gaussian) distribution for the data. For the model of reaction time, we followed the same model selection criteria as for the model of response data and used a model with no correlations between random slopes and intercepts.

Results

Figure S1 shows the influence of prime on reaction time. Participants responded faster after both back and front primes (back prime: 566 ms, front prime: 560 ms) than after neutral primes (644 ms).

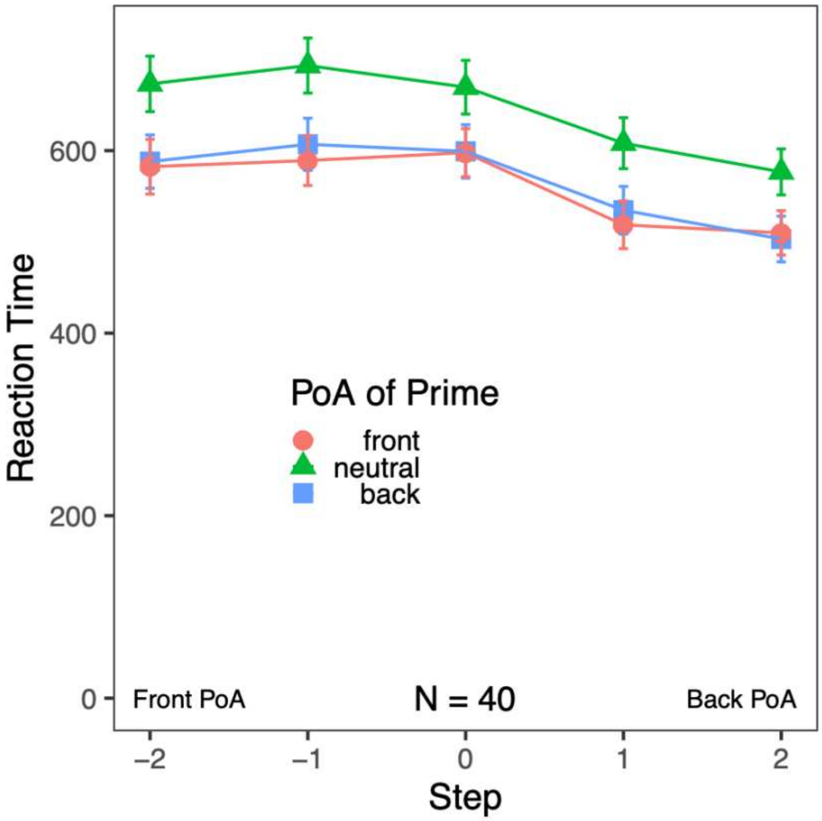

Figure S1. Reaction time for responses in Experiment 1. The x-axis shows the continuum steps, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows the average reaction time, with the colors and shapes indicating prime type.

Our mixed effects logistic regression model revealed a significant effect of Prime (χ^2 = 19.83, *p* < .001), indicating that participants' reaction times were influenced by the written prime. The model also revealed a significant effect of Step (χ^2 = 10.66, *p* = .001) and no interaction.

We ran follow-up tests to analyze pairwise comparisons for Prime, correcting for multiple comparisons as described in the Analysis section. Participants were faster to respond following a front prime than a neutral prime (contrast estimate: 77.1, *z-ratio* = 4.684, *p* < .001), and faster to respond following a back prime than a neutral prime (contrast estimate: 70.6, *z-ratio*

= 6.678, p < .001). However, the difference between reaction time after a front prime and after a back prime was not statistically significant (contrast estimate: 6.45, *z-ratio* = .531, *p* = .848). Thus, these patterns suggest that the presence of any biasing prime made participants faster in their responses.

Experiment 2 Reaction Time Analysis

Response time data were analyzed as above. The maximal model was the best fit, and we used the same linking function as in Experiment 1.

Results

Following what we observed in Experiment 1, we expected to see faster reaction times after participants saw biasing primes (as opposed to neutral primes). Indeed, as shown in Figure S2, participants' reaction times were influenced by the type of prime. Relative to the neutral prime condition, participants responded faster after both back and front primes (also see Table S1).

Table S1.

Reaction time (ms) based on Prime and Context

Context	Prime			
	Front	Neutral	Back	Mean
Back	563	644	571	592
Front	569	653	559	594
Mean	566	649	565	

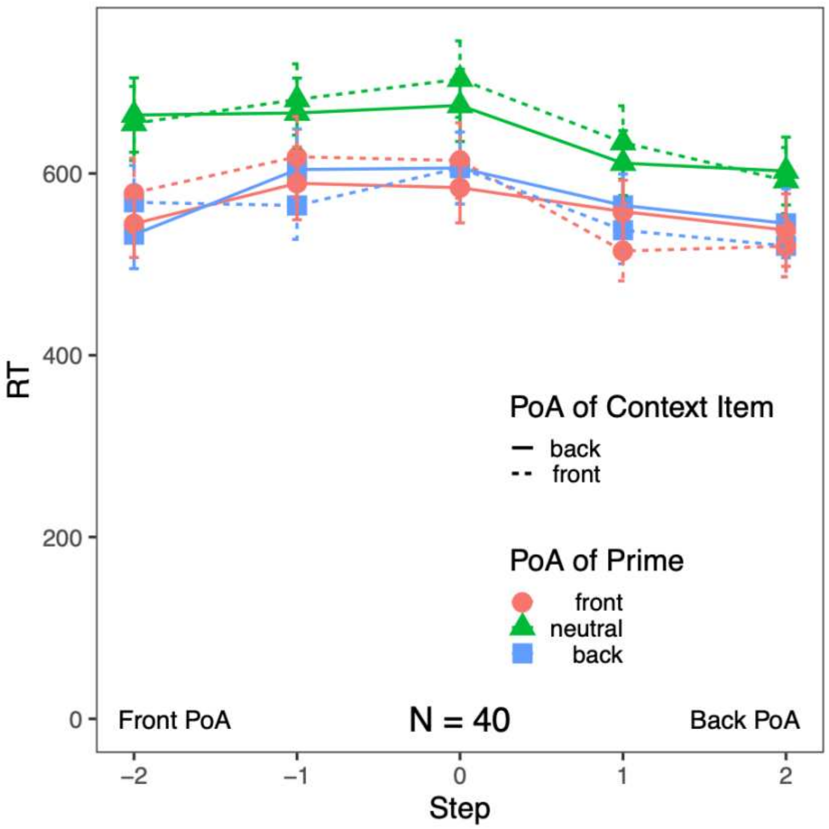

Figure S2. Reaction time for responses in Experiment 2. The x-axis shows the continuum steps, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows the average reaction time. The colors and shapes of the lines indicate prime type. The line type indicates the PoA of the context item.

Our mixed effects logistic regression model revealed a significant effect of Prime (χ^2 = 14.42, *p* < .001), indicating that participants' reaction times were influenced by the written prime. The model also revealed a significant effect of Step (χ^2 = 4.81, *p* = .028) and a significant Step by Context interaction (χ^2 = 25.00, *p* < .001), indicating that the effect of context differed across steps. Figure S3 visualizes this interaction by plotting the raw data collapsed across prime conditions, showing that reaction times change more across step for context items ending in a front PoA than a back PoA.

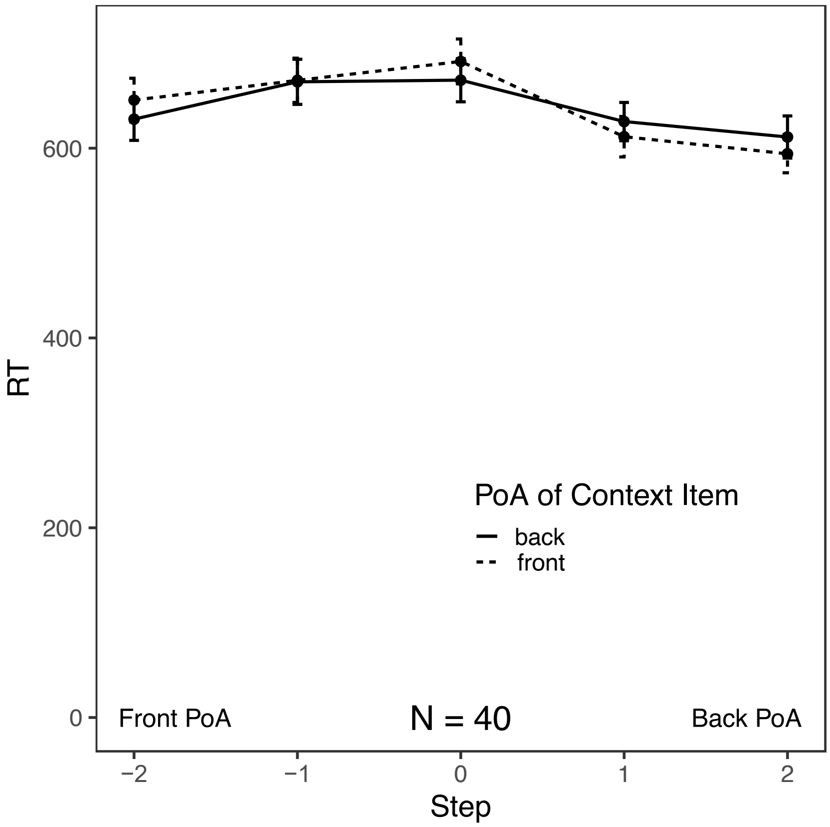

Figure S3. Reaction time based on Context across Step in Experiment 2, collapsed across Prime. The x-axis shows Step, ranging from most front (/s/; -2) to most back (/ʃ/; 2). The y-axis shows the reaction time.

We ran follow-up tests to analyze pairwise comparisons for Prime, correcting for multiple comparisons as described in the Analysis section. Participants were faster to respond following a front prime than a neutral prime (contrast estimate: 83.9, *z-ratio* = 5.677, *p* < .001), and faster to respond following a back prime than a neutral prime (contrast estimate: 84.3, *z-ratio*

= 6.232, *p* < .001). However, the difference in reaction time after a front prime and after a back prime was not statistically significant (contrast estimate: .4, *z-ratio* = .039, *p* = .999).

Experiment 3a Versus Experiment 1 Reaction Time Comparison Analysis

As shown in Figure S4, participants were faster overall for either front primes (average: 525 ms; Experiment 3a: 490 ms, Experiment 1: 560 ms) or back primes (average: 522 ms; Experiment 3a: 478 ms, Experiment 1: 566 ms), as compared to neutral primes (average: 600 ms; Experiment 3a: 556 ms, Experiment 1: 644 ms). Breakdowns of the reaction times by condition can be found in Table S2.

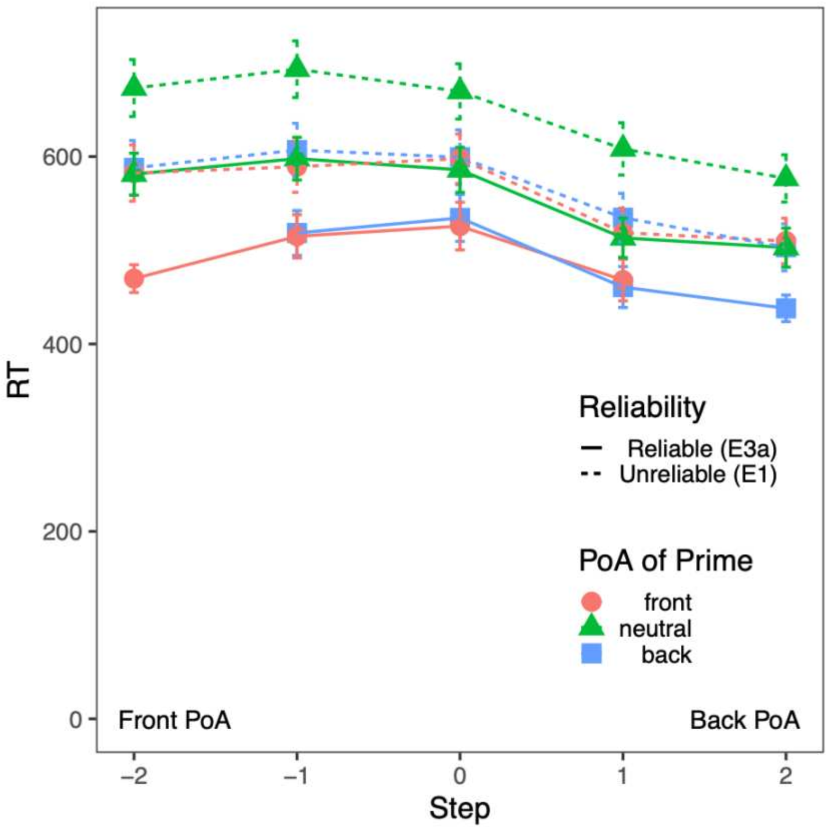

Figure S4. Reaction time data for Experiment 1 (unreliable prime-target relationship) and Experiment 3a (reliable prime-target relationship). The x-axis shows continuum step, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows the raw reaction time, in milliseconds. The colors and shapes indicate the prime that participants saw. The line type indicates the reliability condition. Error bars represent 95% confidence intervals.

Table S2.

Reaction time (ms) by Prime and Reliability.

Reliability	Prime			
	Front	Neutral	Back	Mean
Unreliable (Exp. 1)	560	644	566	590
Reliable (Exp. 3a)	490	556	478	508
Mean	525	600	522	

For our mixed effects model comparing Experiment 1 to Experiment 3a, the model revealed a main effect of Prime (χ^2 = 13.70, *p* = .001), indicating that participants' reaction times were influenced by the written prime. The model also revealed a significant effect of Step (χ^2 = 6.05, *p* = .01). We found no effect of Reliability (reliable mean RT: 508 ms; unreliable mean RT: 590 ms; χ^2 = 1.99, *p* > .05) and no interactions (all p > .05).

Follow-up tests were run to analyze pairwise comparisons for Prime. Participants responded faster after a front prime than after a neutral prime (contrast estimate: 68.8, *z-ratio* = 4.075, *p* < .001), and after a back prime than after a neutral prime (contrast estimate: 62.3, *z-ratio* = 5.733, *p* < .001). However, there were no differences between reaction times after back primes and front primes (contrast estimate: 6.5, *z-ratio* = 0.463, *p* = .885).

Experiment 3b Versus Experiment 2 Reaction Time Comparison Analysis

As shown in Figure S5, participants were faster overall for either front primes (average: 537 ms) or back primes (average: 530 ms), as compared to neutral primes (average: 621 ms). Breakdowns of the reaction times can be found in Table S3.

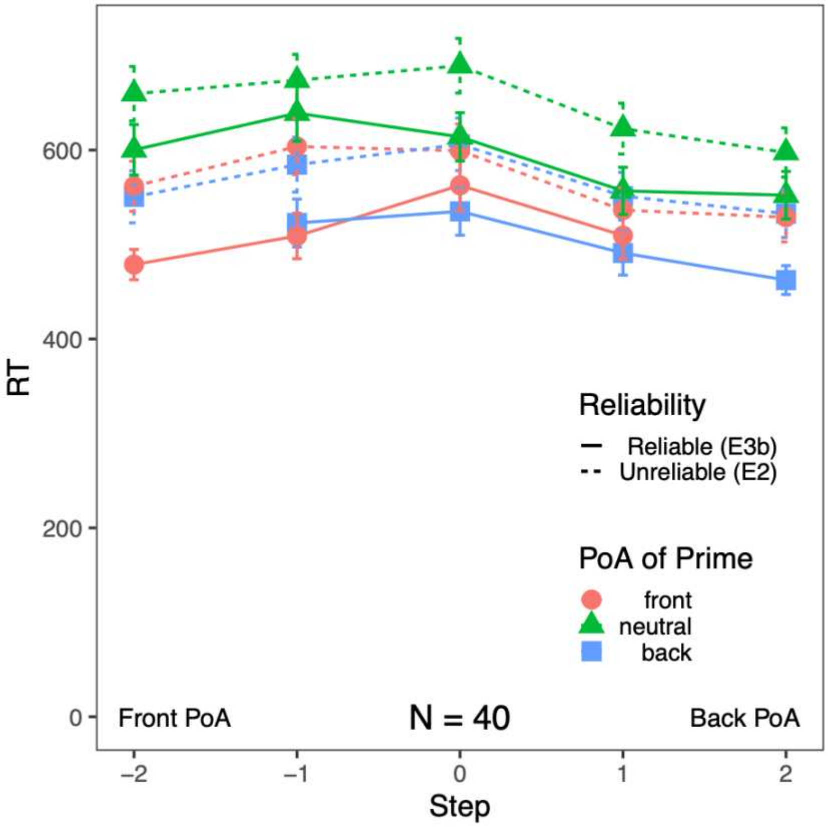

Figure S5. Reaction time data for Experiment 2 (unreliable) and Experiment 3b (reliable) based on prime (collapsed across auditory context). The x-axis shows the continuum steps, ranging from most front (/s/) to most back (/ʃ/). The y-axis shows the raw reaction time, in milliseconds. The colors and shapes indicate the prime that participants saw. The line type indicates the reliability condition. Error bars represent 95% confidence intervals.

Table S3.

Reaction times (ms) based on Reliability and Prime.

Reliability	Prime			
	Front	Neutral	Back	Mean
Unreliable (Exp. 2)	566	649	565	593
Reliable (Exp. 3b)	508	592	495	532
Mean	537	621	530	

For our mixed effects model comparing Experiment 2 to Experiment 3b (which as in the response data model did not include Context to just examine the effect of Prime across experiments), our model revealed a main effect of Prime ($\chi^2 = 28.94$, $p < .001$), indicating that participants' reaction times were influenced by the written prime. We found no effects of Step or Reliability and no interactions (all $p > .05$).

We conducted follow-up tests to analyze pairwise comparisons for Prime. Participants responded faster after a front prime than after a neutral prime (contrast estimate: 71.3, *z-ratio* = 5.800, $p < .001$), and faster after a back prime than after a neutral prime (contrast estimate: 73.2, *z-ratio* = 8.521, $p < .001$). However, there were no differences between reaction times after back primes and front primes (contrast estimate: 1.9, *z-ratio* = .193, $p = .979$).

Individual Differences

To investigate whether there were trade-offs between use of coarticulatory context information and use of the prime, we followed the analysis pipeline from Lai et al. (2022). We note that these additional analyses we did were not preregistered. In order to compare our results to theirs, we used their exact model structure on a subset of our data from Experiment 2, namely, only the non-neutral prime trials. This choice was made in order to mirror their design and analyses, giving us two levels of both prime and coarticulatory context. To understand how individual participants made use of the prime and coarticulatory context, we extracted the by-participant random slopes for prime and coarticulatory context. As shown in Figure S6, individual participants' slopes for prime and coarticulatory context were negatively correlated. A Pearson's correlation test revealed that this correlation was significant (*r* = -0.33, *t*(38) = -2.13, *p* = 0.039), and was similar to the correlation reported by Lai and colleagues (*r* = -0.39).

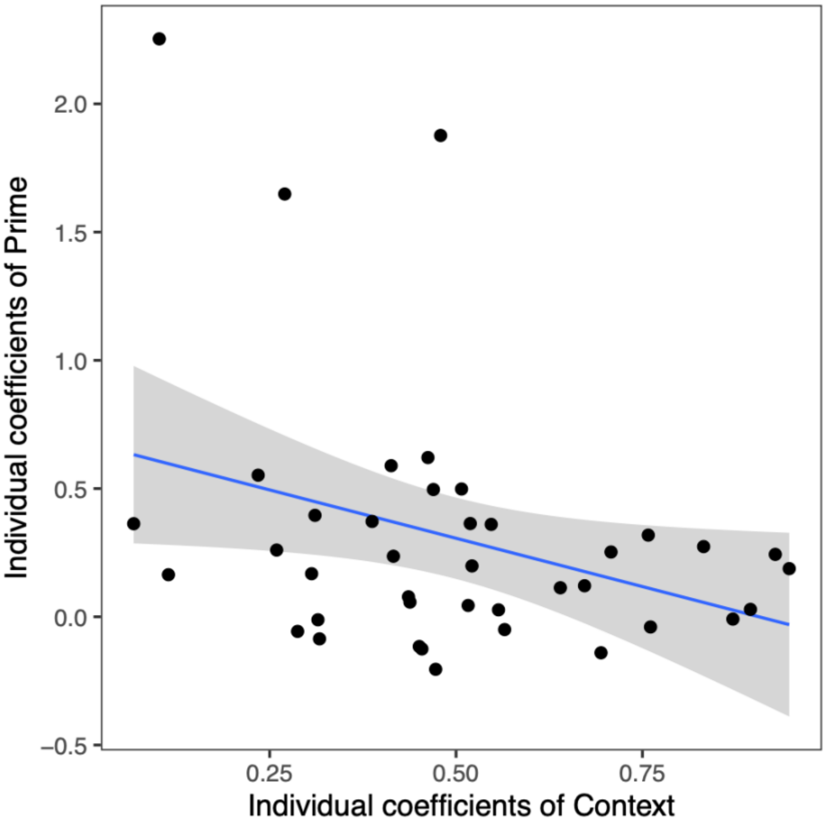

Figure S6. By-participant slope estimates extracted from a mixed effects model for a subset of data (all non-neutral prime trials) from Experiment 2, following the analysis pipeline from Lai et al. (2022). The x-axis shows the estimate for auditory context, and the y-axis shows the estimate for prime. The shaded area represents the 95% confidence interval for the linear regression model between the slope estimates for prime and the slope estimates for auditory context.

These results suggest a tradeoff between use of auditory coarticulatory context and the visual lexical information. In order to test whether this tradeoff exists when we manipulated reliability, we ran the same analysis on data from Experiment 3b, again using just the non-neutral prime trials.

As shown in Figure S7, individual participants' slopes for prime and coarticulatory context were not correlated. A Pearson's correlation test revealed no significant correlation (*r* = 0.12, *t*(38) = 0.74, *p* = 0.465). It is possible, then, that making the prime less reliable led to less of a tradeoff between the two constraints. Another possibility, however, is that because we only analyze the non-endpoint steps for Experiment 3b, the lack of tradeoff could be due to less dominance of one cue for more ambiguous auditory tokens (and conversely, dominance of one cue primarily for endpoint tokens). In order to test these possibilities, we ran the same analysis on only the non-endpoint steps of Experiment 2.

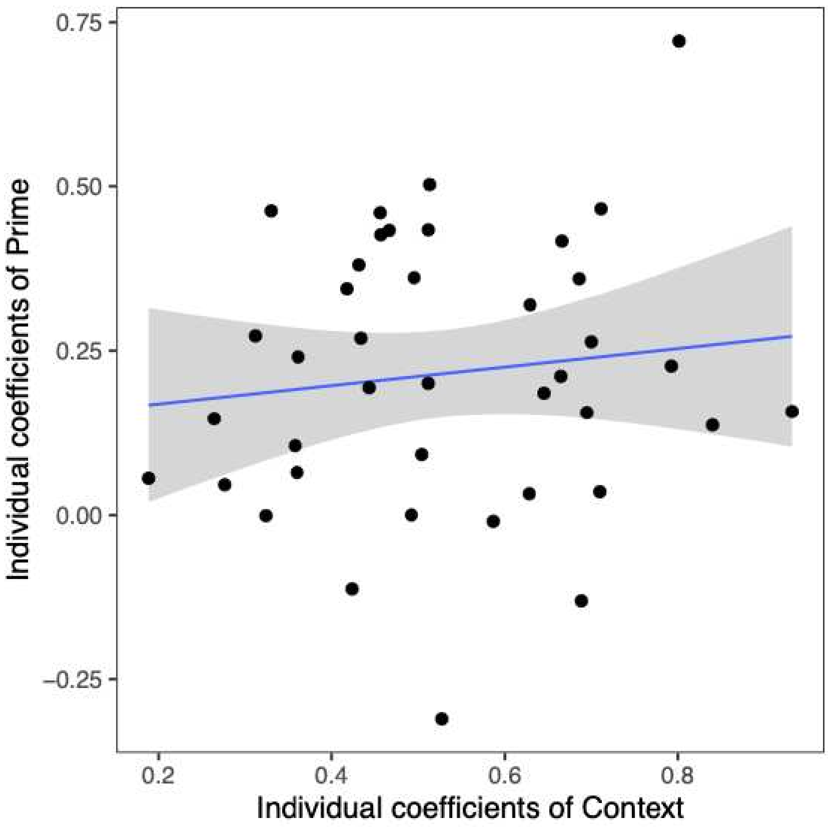

Figure S7. By-participant slope estimates extracted from a mixed effects model for a subset of data (all non-neutral prime trials) from Experiment 3b, following the analysis pipeline from Lai et al. (2022). The x-axis shows the estimate for auditory context, and the y-axis shows the estimate for prime. The shaded area represents the 95% confidence interval for the linear regression model between the slope estimates for prime and the slope estimates for auditory context.

Using only the non-endpoint steps on Experiment 2, the correlation between the by-participant slope estimates for auditory context and prime was not significant (r = -0.10, t(38) = -0.61, p = 0.546). A similar analysis on the middle steps (4 of their 8 continuum steps) from the data from Lai et al. (2022) also reveals a non-significant correlation between the two constraints (r = -0.17, t(40) = -1.12, p = 0.270). Overall then, these results suggest that at middle steps (with more ambiguous acoustics), people may be more likely to be integrating information across multiple constraints, as opposed to more strongly relying on one constraint.