HHS Public Access

Author manuscript

Atten Percept Psychophys. Author manuscript; available in PMC 2020 May 01.

Published in final edited form as:

Atten Percept Psychophys. 2019 May ; 81(4): 1147–1166. doi:10.3758/s13414-019-01712-9.

Unstuck in time: Listeners can anticipate future segments before they identify the current one.

Kayleen E. Schreiber,
Interdisciplinary Graduate Program in Neuroscience, University of Iowa

Bob McMurray
Dept. of Psychological and Brain Sciences, Dept. of Communication Sciences and Disorders, Dept. of Linguistics University of Iowa

Abstract

Speech unfolds rapidly over time, and the information necessary to recognize even a single phoneme may not be available simultaneously. Consequently, listeners must both integrate prior acoustic cues and anticipate future segments. Prior work on stop consonants and vowels (McMurray, Clayards, Tanenhaus, & Aslin, 2008; Reinisch & Sjerps, 2013) suggests that listeners integrate asynchronous cues by partially activating lexical entries as soon as any information is available, and then updating this when later cues arrive. However, a recent study suggests that for the voiceless sibilant fricatives (/s/ and /ʃ/), listeners wait to initiate lexical access until all cues have arrived at the onset of the vowel (Galle, Klein-Packard, Schreiber, & McMurray, in press). Sibilants also contain coarticulatory cues that could be used to anticipate the vowel upcoming. However, given these results, it is unclear if listeners could use them fast enough to speed vowel recognition. The current study examines anticipation by asking when listeners use coarticulatory information in the frication to predict the upcoming vowel. A visual world paradigm experiment found that listeners do not wait: they anticipate the vowel immediately from the onset of the frication, even as they wait several hundred milliseconds to identify the fricative. This finding suggests listeners do not strictly process phonemes in the order that they appear; rather the dynamics of language processing may be largely internal and only loosely coupled to the dynamics of the input.

Keywords

Speech Perception; Spoken Word Recognition; Coarticulation; Anticipation; Cue Integration; Auditory Memory

Corresponding Author Bob McMurray, W311 SSH, Dept. of Psychological and Brain Sciences, University of Iowa, Iowa City, IA, USA 52242 319-335-2408, bob-mcmurray@uiowa.edu.

Open Practices

This study was not preregistered. Materials and processed data are available at https://osf.io/46dex/. Raw data is available by request to the second author.

Introduction

Speech unfolds rapidly in time. Consequently, the acoustic cues for even a single phoneme are often spread across several hundred milliseconds (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Hawkins, 2003; Ohman, 1966; Summerfield, 1981). This creates a problem of integrating information over large temporal regions. Moreover, speech unfolds rapidly, on the order 8–15 phonemes per second, creating a need to access meaning efficiently.

The challenge of accurately identifying words from a temporally unfolding input requires both retrospective and prospective processes. *Retrospectively*, listeners must accumulate information over time and integrate it into phonemic and lexical representations (Marslen-Wilson, 1987; McMurray et al., 2008; McMurray, Tanenhaus, & Aslin, 2009; Warren & Marslen-Wilson, 1987). *Prospectively,* listeners can use fine grained cues in the signal to anticipate future events and speed processing (Gow, 2001; McMurray & Jongman, 2015; Salverda, Kleinschmidt, & Tanenhaus, 2014; Yeni–Komshian & Soli, 1981). Both retrospective and prospective processing are generally seen to conform to a strong view of incrementality that pervades work on spoken language comprehension at all levels, from speech perception to sentence processing (Altmann & Kamide, 1999; Elman, 2009; Frazier, 1987; Gaskell & Marslen-Wilson, 1997; MacDonald, Pearlmutter, & Seidenberg, 1994; McClelland & Elman, 1986). This principle of incrementality argues 1) that listeners accumulate evidence for candidate interpretations as the input arrives, and 2) that they simultaneously use whatever information is available to build evidence for future candidates, even before all the information is available. Thus, both retrospective integration and prospective anticipation are thought to derive from the same basic assumptions about speech perception (and see Gow, 2003; McMurray & Jongman, 2015).

This broader assumption of incrementality is strongly attested throughout psycholinguistics. In speech perception, incremental processing has been observed in the integration of asynchronous cues to a single phoneme (e.g. integrating VOT and the length of the subsequent vowel) (McMurray et al., 2008; J. L. Miller & Dexter, 1988; Reinisch & Sjerps, 2013). Listeners use individual cues to access lexical candidates as soon as they arrive, without waiting for all cues to be available. There is also substantial work in speech perception attesting to listeners' ability to use coarticulatory information to anticipate subsequent segments (Gow, 2001; Sereno, Baum, Marean, & Lieberman, 1987; Yeni–Komshian & Soli, 1981). The few studies that have investigated the timecourse of processing suggest that anticipatory coarticulation is used very early if not immediately (Gow & McMurray, 2007; Salverda et al., 2014), again supporting incremental processing.

However, recent work suggests that listeners may not integrate cues incrementally when identifying sibilant fricatives (Galle et al., in press). Instead, they appear to wait until the onset of the vowel—when most cues have arrived—before they begin to commit to a decision about the fricative. It is unknown when anticipatory processes (e.g., the ability to infer the upcoming vowel from the fricative) are initiated. Do listeners also wait for the vowel to make use of coarticulation? This would appear to undercut the benefit of anticipatory processing, as in this case the listener is only using prior coarticulatory cues

once the vowel (the target they are identifying) is available. The goal of the present study is to examine the precise timecourse of anticipatory processing for fricatives. This is theoretically important, as it will help understand how anticipatory processes relate to retrospective cue integration processes. Indeed, the timecourse of cue-integration and anticipation that we report suggests that the canonical view that speech is processed incrementally may not apply universally to all speech sounds, and raises broader questions about how we should conceptualize the role of time in speech perception.

Retrospective Processing: Incrementally Integrating Asynchronous Cues.

The problem of integrating unfolding material over time is crucial at every level of language; we focus here on the integration of cues to a single phoneme or phonological feature. Consider stop consonant voicing at syllable onset. For this contrast, Voice Onset Time (VOT; the time between the opening of the lips after a stop consonant and the onset of voicing of the following vowel) plays a large role in the percept. Sounds with low VOTs (near 0 ms) are typically heard as /b/'s, /d/'s or /t/'s (in English); higher VOTs (near 50 ms) are heard as /p, t, k/. However, VOT is not the only relevant cue as there are a large number of other cues with somewhat weaker effects (Nearey & Rochet, 1994; Ohde, 1984; Summerfield & Haggard, 1977). For example, in syllable initial position, the length of the subsequent vowel may be used (J. L. Miller & Volaitis, 1989; Summerfield, 1981; Toscano & McMurray, 2012, 2015) and can be seen as a proxy for speaking rate. VOT is a temporal cue, so what "counts" as a low VOT may vary with surrounding speaking rate (which can be indicated in part by vowel length); alternatively VOT may be evaluated relative to the length of the vowel as a sort of contrast effect (Diehl & Walsh, 1989). Crucially, VOT arrives immediately at syllable onset, but the length of the vowel may not arrive until 100 ms later or more.

Obviously, VOT and vowel length cannot be integrated until both arrive at the end of the syllable. But when are they *used*? At least two broad classes of strategies are possible (see McMurray et al., 2008; Reinisch & Sjerps, 2013, for discussion). Under a *late integration* strategy, early information like VOT might be held in a memory buffer and encapsulated from lexical access until vowel length is available. At this point, both are combined and the resulting voicing percept can be used to bias lexical access. In contrast, under an *immediate utilization* strategy early acoustic cues could partially bias lexical activation as soon as they arrive, with later cues further affecting activation. Both are consistent with incremental (left to right) processing, but differ in whether information about individual cues is encapsulated (prior to integration) or available to guide lexical access.

A number of studies examining the integration of asynchronous speech cues support immediate utilization (J. L. Miller & Dexter, 1988; Warren & Marslen-Wilson, 1987). Most recent investigations have used eye-tracking in the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) to assess *when* individual acoustic cues bias lexical competition (McMurray et al., 2008; Reinisch & Sjerps, 2013; Toscano & McMurray, 2012, 2015) or speech perception (Kingston, Levy, Rysling, & Staub, 2016; Mitterer & Reinisch, 2013). These studies generally use at least two cues whose arrival is separated by one hundred ms or more (e.g., VOT and vowel length). Since eye-movements to referents in the VWP reflect lexical-level activation, immediate utilization predicts each

cue will influence the pattern of fixations as soon as the cue arrives. In contrast, late integration predicts that the pattern of fixations will not be influenced by either cue until the second cue arrives and both are available. This suggests the effects of each cue on fixations should be simultaneous and late.

Virtually all of these studies support immediate utilization: early fixations reflect only the first cue and later fixations are affected by the second cue (roughly timelocked to its arrival). This has been observed in stop voicing (McMurray et al., 2008; Toscano & McMurray, 2012, 2015) and manner of articulation (McMurray et al., 2008), and vowels (Reinisch & Sjerps, 2013). Immediate utilization is also seen in cases where more context-driven or top-down factors must be integrated such as sentential speaking rate (Toscano & McMurray, 2015), top-down feedback (Kingston et al., 2016), or prior perceptual learning (Mitterer & Reinisch, 2013).

Immediate utilization is consistent with the incrementality principle that describes many levels of language. More broadly, immediate utilization suggests that levels of language are not encapsulated from each other—early levels of processing (in this case, cue encoding) continuously cascade to higher level ones (in this case word recognition or speech categorization; and see, Apfelbaum, Blumstein, & McMurray, 2011; Levy, Bicknell, Slattery, & Rayner, 2009).

Prospective Processing: Coarticulation and Anticipation.

Work on anticipatory processes in speech offers a close parallel. The acoustic form of any given segment is influenced by surrounding segments (Beddor, Harnsberger, & Lindemann, 2002; Cole, Linebaugh, Munson, & McMurray, 2010; Daniloff & Moll, 1968; Magen, 1997; Ohman, 1966). Consequently, most segments also contain information about upcoming ones. For example, the spectrum of the /s/ in *soup*, has a lower mean frequency than the /s/ in *seep* (Daniloff & Moll, 1968) due to the fact that for *soup,* speakers round their lips anticipating the rounding in the vowel. In principle, such information could be used to identify the phonological vowel earlier than if the listener had to wait to the onset of the voiced portion.

Many perceptual studies have focused on fricative-vowel sequences such as /si/ or /ʃu/. This is a useful paradigm in which to study anticipation as the fricative portion can be clearly isolated from the following vocalic portion or *vocoid*. We use the term vocoid here to refer to the section of the signal containing periodic voicing. This term is used to refer to the voiced segment as a whole, distinct from the phonological content of this portion (e.g., the specific vowel), and makes no assumption about what constitutes a vowel.

Studies of vowel anticipation suggest that listeners can use coarticulatory information in the frication to “guess” upcoming vowels at above chance levels (McMurray & Jongman, 2015; Nittrouer & Whalen, 1989; Sereno et al., 1987; Yeni–Komshian & Soli, 1981) (and see Jenkins, Strange, & Edman, 1983; Parker & Diehl, 1984, for parallel work on coarticulation in vowels). Work with other cues also suggest that listeners are faster to identify upcoming segments when the coarticulatory information in the preceding vowel matches the identity of those segments (Gow, 2001; Martin & Bunnell, 1981). Such studies document that listeners

can hear and use this fine grained coarticulatory information that precedes the primary cues for the target segment.

These studies leave open however the timecourse of this process and could be consistent with either the immediate utilization or late integration accounts described for cue integration. First, under an immediate utilization (or incremental processing) account, coarticulatory modifications in the input could build activation for the target segment (the vowel) before it arrives. However, an alternative is that listeners wait to activate the target segment until it has arrived (e.g., they begin activating a vowel at the vocoid onset). At this point, they can activate it more robustly or more efficiently when the preceding coarticulation matches[1]. These hypotheses are quite similar to those posited in retrospective cue integration—indeed the only real difference may be the phonetic phenomenon to which they are applied.

Eye-tracking in the VWP can be used to distinguish these hypotheses by measuring *when* coarticulatory information is used. Several studies have manipulated coarticulation in a preceding sound and used the VWP to isolate when coarticulation is used (Gow & McMurray, 2007; Salverda et al., 2014). These studies support early utilization: fixations to lexical competitors are biased by coarticulatory information before the target segment arrives.

Fricatives.

The work described thus far suggests that both integration and anticipation begin as soon as information is available, and this cascades as far as lexical-level processing. At the broadest levels, these findings of immediate utilization (for both processes) are consistent with incremental (left-to-right) processing: listeners make partial inferences about both current and future segments and update these as further information arrives.

However, recent studies suggest that in fricatives, retrospective cue integration may work differently. Isolated fricatives in syllable initial position (e.g., those that are not in clusters), are typically characterized by 100–200 ms of high frequency turbulence, followed by a periodic vocalic portion or vocoid. Again, we use the term vocoid here to refer to the vocalic chunk of signal, independently of the identity of any specific vowel)[2]. In fricatives, place of articulation is cued both by the frequency spectrum of the frication, and formant transitions in the vocoid (among many other cues, Jongman, Wayland, & Wong, 2000). This creates a long temporal window over which cues must be integrated. In sibilant fricatives, the frication typically comprises information above 3000 Hz, while the most important information in the vocoid (e.g., the formant transitions for identifying the fricative, or the formants for identifying the vowel) lies largely below 3000 Hz. Consequently, if speech perception follows from fundamental auditory principles (Diehl, Lotto, & Holt, 2004; Holt & Lotto, 2008), the large acoustic disparity between the cues for fricative identity that are contained

[1]This latter hypothesis may appear to be ruled out by studies using gated stimuli in which the vowel never arrives; however, listeners may adopt a different approach to gated stimuli since they know further information is not forthcoming.
[2]This term makes no assumptions about when a phonological vowel begins or ends, the vocoid merely marks the portion of the signal where modal voicing occurs.

in the frication and those contained in the vocoid (e.g., the formant transitions) may make it more challenging to integrate cues over time.

Two studies report data relevant to the timecourse of fricative perception (Kingston et al., 2016; Mitterer & Reinisch, 2013). Both examined fricatives (the /s/ vs. /f/ contrast), but this was not their primary goal as they were concerned specifically with the use of higher level lexical factors and perceptual learning (not fricative identification). While neither set out to test claims about how bottom-up cues were integrated, both show fairly early use of cues in the frication. However, in both studies, several methodological factors prevent strong claims about the integration of fricative cues (see, Galle et al., in press, for a longer discussion). For example, both studies used letters (not pictures) as response cues, preventing clear claims about when lexical access begins. Moreover, neither study manipulated a second acoustic cue to compare the timing of the utilization of frication to other cues or to the vocoid. Finally, Mitterer and Reinisch (2013) used word-final fricatives, where the relevant vocoid precedes the frication).

We recently examined retroactive cue integration in the context of sibilant fricatives (Galle et al., in press) using a design more similar to (McMurray et al., 2008) to answer this question more definitively. Listeners heard continua spanning two sibilants (e.g., *ship/sip, shoot/suit*) which varied orthogonally in two dimensions: the frication spectrum (which varied in several steps between /s/ and /ʃ/, and the formant transitions in the vocoid (which were either consistent with an /s/ or /ʃ/). This study also varied a third factor, the rounding on the following vowel, for a 7×2×2 design. Sibilants produced in front of a rounded vowel like /u/ tend to have a lower spectrum than in front of an unrounded vowel like /i/ (Daniloff & Moll, 1968). Consequently, fricative identification can be improved if listeners know the vowel rounding and use it to shift their percept of the fricative (Apfelbaum, Bullock-Rest, Rhone, Jongman, & McMurray, 2014; Mann & Repp, 1980). In the domain of fricative identification, rounding is a context effect in that the rounding does not directly reveal what fricative it is, but it allows listeners to compensate for this coarticulation.

Galle et al. (in press) used the VWP to ask when listeners use the frication spectrum, the formant transitions in the vocoid, and vowel rounding to bias lexical competition between /s/ and /ʃ/-initial words. Surprisingly, the effect of the frication spectrum on fixations was extremely late, and coincident with the effects of both formant transitions and rounding. Listeners appeared to wait to identify the fricative for several hundred milliseconds, until all three sources of information were available. This was replicated in four experiments. First, these experiments showed that this effect did not derive from methodological factors such as the number of factors manipulated in the experiment or the style of stimulus construction (it held with unmodified recordings). Second, they showed that listeners were not delaying judgement because of insufficient information: an experiment with gated stimuli showed there was sufficient information in even the first 50 msec of the frication for accurate identification. Third, late integration held even in running speech; this rules out the hypothesis that listeners need access to the talker (which normally cannot be identified until the vocoid) as this would have been available in the preceding sentence. Fourth, manipulating the length of the fricative changed the timing of when the fricative cues are used—when fricatives were longer, the onset of the frication effect was delayed. This clearly

documents that listeners are waiting for the onset of the vocoid (or the end of the fricative). Finally, an analysis across experiments showed that during the several hundred milliseconds of the frication, not only were listeners not committing to one of the fricatives (/s/ vs. /ʃ/), but they were not even committing differentially to fricatives over stops until vocoid onset—they didn't even know that the stimulus was a fricative!

It is not clear yet whether this finding applies beyond voiceless sibilants (/s/ vs. /ʃ/); however, this work clearly identifies that not all speech sounds are processed incrementally—some can employ late integration. This unique case is leveraged here to understand the relationship of anticipation and cue integration.

Anticipation in Fricatives.

One immediately obvious hypothesis for the late integration observed in /s/ and /ʃ/ is that this effect is due to the acoustic properties of fricatives. To perhaps a greater extent than other sounds, information in fricatives is spread across highly distinct frequency bands. If speech is perceived in a way that is consistent with general auditory principles (Diehl et al., 2004; Holt & Lotto, 2008), classic work in auditory streaming (Bregman, 1990) suggests this kind of separation may make it difficult to perceive a unified auditory object.

When we consider the implications of an auditory account for anticipatory processes, this makes a clear prediction. Consider a situation in which listeners may use coarticulation in the sibilant to predict the subsequent vowel (Daniloff & Moll, 1968; Yeni–Komshian & Soli, 1981). Here again we have the same issue: vowels have the majority of their energy in lower frequency regions of the acoustic space, while the coarticulatory cues are in the much higher frequency frication (see Supplement S3 for a phonetic analysis documenting this in our stimuli). This then predicts that the use of this anticipatory information may not take place until the vocoid arrives. This is perhaps too late to be truly anticipatory, though it may lead to a more robust percept.

A second possibility, however, is that late integration is not a product of the acoustic properties of the input, but rather from the phonological unit being identified. Diehl, Kluender, Foss, Parker, and Gernsbacher (1987) present data showing that listeners' reaction times to identify place of articulation in syllable-initial consonants (/b/ vs. /d/ vs /g/) is influenced by the inherent duration of the *following* vowel. They propose that vowels may serve as perceptual anchors or organizers for the syllable; that is hearing or identifying the vowel may be necessary for perceiving other phonemes within the syllable. This is somewhat at odds with the (now) predominant view of incrementality; however it is consistent with the older "P-center" hypothesis (Hoequist, 1983; Marcus, 1981). To adapt such an account to this particular problem we might assume that it is not the center of the vowel that is necessary for anchoring the percept of the syllable, rather listeners simply need enough information to begin identifying the vowel (e.g., the onset of the vocoid). If listeners must begin identifying the vowel before they can work on the rest of the phonemes in the syllable, this could explain the late integration of fricatives. Listeners must wait for the onset of the vowel to begin making distinctions among consonants, so they delay any work identifying the fricative until the vowel has arrived. It can also explain immediate utilization for stop consonants and approximants (and obviously vowels). In these sounds the vowel can

be deduced extremely early and accurately from coarticulatory information. For example, work on silent center vowels (Jenkins et al., 1983; Jenkins, Strange, & Miranda, 1994; Parker & Diehl, 1984), suggests that listeners can identify vowels from formant transitions alone as accurately as they can from unbroken vowels.

This vowel-focused account makes a counterintuitive prediction for anticipatory processes. If the vowel is so important, listeners may begin identifying the vowel (on the basis of coarticulatory information) during the frication, even as they wait for the vowel to identify the fricative. Under this view, they may essentially identify the second phoneme first, and the first phoneme second. This view receives loose support from a gating study by Nittrouer and Whalen (1989) which showed that at some early gates, listeners may be able to reliably deduce the vowel even as they cannot identify the fricative.

While several studies document that people can use anticipatory information during fricatives to identify the vowel (McMurray & Jongman, 2015; Nittrouer & Whalen, 1989; Sereno et al., 1987; Yeni–Komshian & Soli, 1981), these studies all use phoneme identification measures in which people simply report which vowel they think is coming. As a result, they have not examined the timecourse of when this information is used. Similarly, fine grained analyses of the timecourse of processing in fricatives (Galle et al., in press) have only examined fricative identification, not vowel anticipation; and fine grained analyses of the timecourse of anticipatory processing (Gow & McMurray, 2007; Salverda et al., 2014) have not examined the unique case of fricatives. Thus, this experiment asks when coarticulatory information during the frication is used to identify the subsequent vowel.

The Present Study.

Following prior work on retrospective integration (Galle et al., in press; McMurray et al., 2008) and anticipation (Salverda et al., 2014), we used eye-tracking in the VWP to ask when different sources of information are used during speech perception. We focused on the sibilant fricatives /s/ and /ʃ/, and coarticulation for lip rounding (e.g., *see* vs. *sue*) a particularly strong form of anticipatory coarticulation. While Galle et al. (in press) extensively examined the problem of retrospectively integrating cues for identifying the fricative, we focus here on the converse: using cues in the frication to identify the vowel. That is, we ask when listeners can use coarticulatory information to identify the vowel (which was not addressed by prior work).

Our study used a set of close minimal pairs to identify when listeners identified both the fricative and the vowel. First, to determine when listeners could identify the fricative, we examined fixations on trials when the contrasting pictures represented two distinct sibilants with the same vowel (such as *seed* vs. *sheep*). Similarly, to determine when they could identify the vowel, we examined fixations to *seed* vs. *soup.* Note that on any trial the words were always disambiguated by the final phoneme (given the particular response options available on that trial), so the stimuli were eventually unambiguous. Thus, our analyses focused on the eye-movements leading up to the decision.

The auditory stimuli were manipulated in three ways to ask *when* the contents of the stimulus influenced these two decisions. First, to examine the fricative cues (in the frication)

we used natural recordings of words that started with each fricative (e.g., *seed* and *sheep*). By comparing fixations to each word as a function of the auditory stimulus we could determine the time at which fricative decisions were made. Second, to determine when the formant cues in the vocoid were used to identify the vowel, we compared fixations to the vowel targets (*seed* vs. *soup*) between natural recordings of the two vowels (we controlled for coarticulation in the fricative as we describe next). Third, we used the filler items (/b/ and /p/-initial words) in a similar way to identify when listeners could make a stop distinction. These three sets of onsets allow a replication and extension of Galle et al. (in press). We replicated it by testing the hypothesis that if fricatives are integrated late, we should expect to see a later effect of frication than of stop consonant voicing. However, we also extend it by comparing the onset of the frication effect to that of the vowel: our prior work only examined when people identify the fricative (showing that the frication and vocoid cues are used simultaneously) – here we extend this by asking if fricatives and vowels are identified simultaneously.

Finally, and most importantly, we examined the use of coarticulation in the frication to make vowel decisions. To do this, we cross-spliced fricatives across vowels (e.g., the /s/ from *soup* [containing coarticulatory cues indicating rounding], was spliced onto the *eed* from *seed*). We compared conditions where the coarticulation matched or mismatched the vowel (e.g., s_i*eed vs.* s_u*eed* and s_u*oup* vs. s_i*oup*). Here, an anticipatory effect should appear as increased looks to the item whose vowel matched the coarticulation in the frication. Critically, we compare the timing of the anticipatory effect to that of the fricative effect. A simultaneous effect would suggest a purely auditory account—listeners simply can't use any information from high frequency frication until late. In contrast, if the coarticulation effect precedes the fricative effect, it would support the vowel-centered account.

Methods

Participants.

Thirty-one monolingual English speakers from Iowa City (17 female) were recruited in accordance with university human subject protocols. Participants were undergraduates at the University of Iowa and between 18 and 22 years of age. They received $15/hour or class credit for their participation. Participants self-reported English as their only language, normal hearing and normal or corrected-to-normal vision.

Design.

Typical studies on the timing of cue integration (Galle et al., in press) use continua that vary along one or more acoustic dimensions (e.g., the frication spectra). However, many of the same acoustic properties that would be manipulated to create an /s/→/ʃ/ continuum would also be affected by rounding coarticulation. For example, /s/ in a rounded vowel context has a lower spectral mean, a cue which is also one of the most prominent cues to fricative place of articulation (Forrest, Weismer, Milenkovic, & Dougall, 1988; McMurray & Jongman, 2011). Thus, to avoid potential confounds, we used natural recordings with a complete phoneme difference (/s/ vs/ /ʃ/, /i/ vs. /u/) to manipulate the frication and vowel, and used cross-splicing to manipulate coarticulation. This use of mostly intact stimuli (rather than a

continuum) is not unprecedented for studying this issue: Experiment 3 in Galle et al. (in press) used this approach to confirm late integration of fricatives.

Item sets were constructed to contrast both the fricative and following vowel. There are few four-way minimal sets in English (e.g., *suit, shoot, seat, sheet*), so we opted instead to use words that would temporarily compete, but be disambiguated by the final phoneme (e.g., words that share a fricative and vowel, but not the final consonant). As a result, our measure focused on bias in the eye-movements during the ambiguous period. We constructed sets of words in which individual pairs could express each contrast (fricative or vowel), but for which across pairs we could assess both contrasts. For example, in one set, the fricative contrast could be examined by comparing fixations to *seed* and *sheep* (in the context of an unrounded vowel) or between *soup* and *shoot* (in the context of a rounded vowel). Within that same set, we could evaluate vowel identification by contrasting *seed* vs. *soup* (in the context of an /s/), and *sheep* vs. *shoot* (in the context of an /ʃ/).

In the VWP, phonologically unrelated words are typically used as a baseline. However, with four possible items, this would require five items on the screen, exceeding the capacity of visual working memory (4 items: Luck & Vogel, 1997). Consequently, on any trial all four items in a set were not presented, but only two that highlighted a particular contrast. These were paired with two phonologically unrelated words (/b/- or /p/-initial words). Thus, each item-set included four fricative-initial items (/s/ vs. /ʃ/ × rounded vs. unrounded vowels) and four stop-initial items (two /b/'s, two /p/'s), and each trial used two experimental and two filler items.

Within each item-set, we created four groupings of four words for use on a given trial (Table 1). Each grouping featured one pair from the set, selected to highlight one contrast (e.g., /s/ vs. /ʃ/), in one context (e.g., in unrounded vowels, *sheep* vs. *seed*). Groupings were defined by which items contrasted on the screen relative to the auditory stimulus. All four words within a given grouping (including the two fillers) were heard across trials. Given this structure, each word did not always appear with the same competitors on the screen, but each word was equally likely to be the target, competitor or unrelated item across trials.

Vowel groupings contained words that matched on the initial fricative, but differed on the rounding of the vowel. By examining relative fixations to the rounded or unrounded item on these trials, we assessed when the formants in the vocoid and when the coarticulation in the fricative distinguish rounded and unrounded vowels. Across groupings, vowel trials included words starting with both fricatives (though the specific fricative was always the same within a trial).

Fricative groupings contained words that had the same vowel but differed in their initial fricatives. These examined the degree to which the frication spectrum distinguished /s/ from /ʃ/. These included words with both rounded and unrounded vowels.

Across sets, a range of vowels were employed, with rounded vowels consisting of /oʊ/, /u/, /ɔ/ and /ʊ/; and unrounded vowels consisting of /i/, /ɪ/ and /ɛ/. Filler items always consisted of /b/ or /p/ competitors that were not minimal pairs. Lexical statistics (neighborhood density and frequency) and phonotactic probabilities are reported in

Supplements S1 and S2. They document that 1) there are not lexical confounds in these items; 2) the items are within a reasonable range for other similar experiments; and 3) /s/ and /ʃ/ do not have different phonotactic biases with respect to vowel rounding. Even if there had been differences, however, we note that these lexical and phonotactic factors are not an issue as the relevant contrasts are within item (e.g., items appear with both rounded and unrounded coarticulation) and counterbalanced (both fricatives are tested with both types of vowels and in both counterbalancing conditions).

The onset consonants of the fricatives were cross-spliced onto the vocoids to manipulate coarticulation. The formant transitions (in the vocoid) were always appropriate for the onset consonants, but the coarticulation in the frication could predict a rounded (e.g. /s/ from *soup*) or unrounded (/s/ from *seed*) vowel. In order to avoid drawing attention to the fricatives, we also cross-spliced small portions of the filler items. We conducted an extensive phonetic analysis (Supplement S3) which documents that coarticulatory cues are present for both fricatives, can be seen in the high frequencies, and are present as early as the first 50 msec.

There were six total word sets (see Appendix). These were crossed with 4 groupings/set × 4 targets/grouping × 2 splice-conditions for 192 total test stimuli. Each of these was repeated three times, for 576 trials. These were completed in a single hour-long session.

Auditory Stimuli.

A male talker recorded six exemplars of each word in a neutral carrier sentence (*He said…*"). Recordings underwent noise-reduction by estimating the background noise spectrum from a silent period and filtering it from the stimuli using Audacity (Audacity Team, 2015). The first three exemplars were then chosen and excised from their carriers.

For the fricatives, the frication was cut from the vocoid at the 0-crossing nearest the onset of the vocoid. This was defined as the place where high frequency oscillations were no longer visible in a zoomed in view of the wave form. The frication portions were then spliced onto the vocoids. This was done for each of the three exemplars to create six versions of each word (three exemplars × each type of fricative coarticulation). For matching coarticulation, the fricative and vocoid never came from the same exemplar. The average length of frication was 264 ms (not including the 100 ms of silence at stimulus onset).

For b/p stimuli, any prevoicing was eliminated from the original recordings (as this is unnecessary for identifying voicing in English). The onsets were cut and spliced onto separate vocoids so that all stimuli had been spliced. For the /b/-initial items, this was done at the zero-crossing closest to 30 ms. For the /p/-initial items, we cross-spliced the entire aspirated portion.

Finally, all stimuli were amplitude normalized with Praat (Boersma & Weenink, 2009).

Visual Stimuli.

Visual stimuli were clipart-style images of each of word, constructed using a standard lab protocol. For each item, several pictures were downloaded from a commercial clipart database. A group of students then selected the most canonical exemplar for a word, and the

selected pictures were edited for consistent color and brightness, to eliminate distracting elements, and to ensure that each image was a highly prototypical representation. Finally, each picture was approved by a senior member of the research team with extensive experience with the VWP.

Procedure.

The experiment was run using Experiment Builder (SR Research Ltd., Ottawa, CA). Participants were seated in front of a PC with a 19-in. CRT monitor in a sound attenuated, dimly lit room. The eye-tracker was calibrated using a standard nine-point calibration. Auditory stimuli were presented over Sennheiser HD 280 headphones amplified by a Samson C-que 8 amplifier. Throughout the experiment, participants could adjust the volume to a comfortable level.

Prior to testing, participants were familiarized with the pictures. Each picture was shown with its name printed below it. The participant studied each picture/name pair and pressed the spacebar to advance. Next, the participant received written instructions.

During the experimental phase, participants saw four pictures, one near each corner of the screen. Pictures were 300×300 pixels (6.4 ° at a viewing distance of 50 cm), and separated by 780 pixels (24.5°) horizontally and 524 pixels (16.6°) vertically. On each trial, pictures for two members of a fricative pair were present along with two members of a stop-consonant pair (according to the condition). Pictures were randomly assigned to the four locations on each trial using a constrained randomization algorithm that ensured that each of the 24 possible arrangements of pictures were roughly equally likely.

At trial onset, a blue dot (50-pixel diameter) was displayed in the center of the screen along with the pictures. After 500 ms, the dot turned red, and the participants clicked on it to hear the auditory stimulus. This short pre-scan ensured that the participant knew the object locations prior to the stimulus (minimizing the need for visual search). It also ensured that the mouse (and likely the gaze) were centered at trial onset. After hearing the stimulus, participants clicked the picture that best represented the auditory word. After the participant clicked on a picture, the display disappeared and 500 ms later, the next trial began. Participants were encouraged to take their time and perform accurately.

Eye Movement Recording and Analysis.—Fixations were recorded with an SR Research Head Mounted eye-tracker operating at 250 hz. Where possible, both eyes were tracked, but only the eye with the better calibration was used. The eye-movement record was parsed into saccades, fixations, and blinks using the default “psychophysical” parameters. Each fixation was grouped with the preceding saccade into a single unit (a look), which started at saccade onset, ended at fixation offset, and was directed to the location of the fixation. In matching fixations to objects, 100 pixels were added to the border of each object (in each side) to account for noise in the tracking. This did not result in any overlap between regions of interest. A drift correction procedure was conducted every 20 trials. Time was coded relative to trial onset (100 msec before stimulus onset). It is known to take approximately 200 msec to plan and launch an eye-movement (Viviani, 1990); time was not

adjusted by this factor. Thus, an event in the fixation record at 300 msec should be interpreted as planned at 100 msec.

Results

Overview

Average accuracy was 99.36%. As expected, our minimally manipulated stimuli were easily recognized, and no statistics were conducted on accuracy.

To examine the timecourse of processing, we started by computing the proportion of trials on which participants fixated each object at each 4 ms time step (Figure 1). Here, time is coded relative to trial onset, with the stimulus appearing 100 msec later. Vertical lines indicate frication onset and mean vocoid onset after adjusting for the 200 ms oculomotor delay.

Figure 1A and B show the results for trials using a fricative grouping. Fixations to the target start to diverge from those to all other objects a little before 500 ms, but there's a marked acceleration of this trend at around 600 ms. Notably, the object depicting the competing fricative (e.g., *seed* when the target was *sheep*) does not appear to receive any consideration: there were no more fixations to the competing fricative than to a phonologically unrelated word (e.g., *bean*). This is consistent with the idea that by waiting longer to make the judgement, listeners can be more certain and show less competition (McMurray, Farris-Trimble, & Rigler, 2017).

Figure 1C and D show a similar analysis for vowel groupings. Here, the two vowels start to separate from each other at around 500 ms. However, in contrast to the fricatives, there is some consideration for the competing vowel (e.g., *soup* when the target is *seed*).

To examine the effect of coarticulation, Figure 2 shows data from the vowel groupings (Figure 1C, D) as a function of the coarticulation in the fricative (averaged across both /s/ and /ʃ/ fricatives. Figure 2A, for example, shows fixations when the stimulus included a rounded vowel. Here, the fixations to the rounded item (black lines) deviated from the unrounded (gray lines) earlier when the frication contained matching (rounded) coarticulation (solid lines) than non-matching (unrounded) coarticulation (dashed lines). Figure 2B shows the reverse pattern when the stimulus contained an unrounded vowel. In both cases, the effect of coarticulation (the difference between the dashed and solid lines) appears to begin early – during the frication – and disappear by around 1000 ms.

Contrasts and their Timing.

The original analysis plan called for an analysis that asked when each factor (the frication, the vowel, and the coarticulation in the frication) affected ongoing decisions in the fixation record. For this, we used an onset detection technique developed in prior studies (McMurray et al., 2008; Reinisch & Sjerps, 2013) and originally developed for event related potentials (J. O. Miller, Patterson, & Ulrich, 1998).

We started by converting the fixations at each 4 ms time bin to a measure of bias, the difference between fixations to a target and competitor object (Clayards et al., 2008; Galle et al., in press). This collapses fixations to any one of four items into a single measure of how well the listener is contrasting fricatives or vowels at that moment. For the fricative groupings, we computed $Bias_{ʃs}$ as the proportion of looks to the /s/ object minus the proportion to the /ʃ/ objects. This is near 1.0 if the participant was fixating /s/, −1.0 if the participant was looking at /ʃ/, and 0 if they were equi-biased (or not looking at anything). For the Vowel groupings, $Bias_{unround}$ was the proportion of looks to the unrounded vowel word minus the proportion to the rounded word. We also analyzed the filler trials (/b/ vs. /p/) by computing $bias_{bp}$, the proportion of looks to the /b/-initial item minus the proportion to the /p/-initial item.

From these bias measures we then computed several contrasts, which we term “effects”, as a function of time. First, we estimated how strongly listeners used the frication spectrum to distinguish /s/ from /ʃ/ on trials with a fricative grouping (the *fricative effect*). This was done by averaging $bias_{ʃs}$ across both rounding conditions but within each fricative /s/ vs. /ʃ/. We then subtracted $bias_{ʃs}$ on /ʃ/ trials from that on /s/ trials, to determine how strongly the frication drives bias toward the target object.

Second, we asked how strongly information in the vocoid (presumably formant frequencies) is used to distinguish vowels on trials with a vowel grouping (the *vowel effect*). This was done by examining $Bias_{unround}$ (looks to the unrounded vowel [e.g., *seed*] minus the rounded vowel [e.g., *suit*]) as a function of the vocoid, regardless of the coarticulation on the fricative (which was varied orthogonally). That is, we contrasted $bias_{unround}$ between stimuli with rounded (s_i*uit* and s_u*uit*) and unrounded (s_i*eed* and s_u*eed*) vowels.

Third, we asked when listeners used coarticulation in the fricative (e.g., s_u*oup* vs. s_i*oup*) to distinguish rounded from unrounded vowels on trials with a vowel grouping (the *coarticulation effect*). Thus, we subtracted $bias_{unround}$ on trials with rounded coarticulation in the fricative from those with unrounded fricatives. This was expected to have only a short-lived effect (as it would be superseded by the unambiguous vocoid when it arrives later).

Finally, on the control (/b/ vs. /p/) trials, we asked how strongly listeners distinguished the two stop consonant initial words (the *stop effect*). We note that this latter contrast also includes a vowel contrast (e.g., *bean* and *pail* differ in both voicing and vowel). As a result, this contrast was not of theoretical interest (as it was in McMurray et al., 2008; Toscano & McMurray, 2012), but was intended identify the earliest moment when the fixation record might exhibit a large stimulus-driven difference.

Each of these four effects were computed for each subject at each 4 msec timeslice. We then made several planned comparisons to test targeted hypotheses:

- We compared the timecourse of the stop and fricative effects to replicate Galle et al. (in press) who showed that fricative perception was highly delayed relative to stops.
- We compared the timecourse of the stop and fricative effects to extend Galle et al. (in press). That study showed that the timing of the frication effect did not

differ from cues to the fricative at vocoid onset (e.g., the formant transitions), but they did not compare it to the identification of the vowel itself.

- The most important novel contrasts in this study were between the timecourse of the coarticulation and both the fricative and vowels effects. We asked if listeners must also wait for the vocoid before they can use the coarticulation (predicting late simultaneous effects), or if they can use coarticulation immediately even as they must wait for the vowel.

Magnitude of effects

Our primary analysis asked *when* each effect occurs (e.g., when the effect of coarticulation is seen). However it was important first, to validate that each effect can be observed, before assessing its timing. Our planned approach—based on prior work (Galle et al., in press; Toscano & McMurray, 2015)—was to use an ANOVA or mixed model to examine area under the curve for targeted regions. However, there is increasing interest in approaches which simultaneously test a hypothesis and identify a timewindow in which it is significant (Maris & Oostenveld, 2007; Oleson, Cavanaugh, McMurray, & Brown, 2017; Seedorff, Oleson, & McMurray, in press). Such techniques do not statistically compare the onsets of effects; rather they identify the rough region in which an effect is operative. Moreover, the precise timing of these effects is a function of both the absolute difference between conditions, and the variability. As a result, an effect could appear later (in one condition or another) due to differences in the variability, not differences in when the effect appears. Thus, these analyses are not appropriate for our primary question, and we reserve inference on the difference among conditions for the planned onset detection analysis which was designed to answer this question. However, they offer 1) a way to identify *that* an effect was present without making assumptions about the time window of interest; and 2) a useful exploratory tool for highlighting potential time regions. As these were not the analytic approach that was planned for this study—they were suggested by an anonymous reviewer—they must be treated as exploratory.

While both the Bootstrapped Difference of Timeseries (BDOTs; Oleson et al., 2017) and Cluster Based Permutation analyses (Maris & Oostenveld, 2007) offer ways to do this, we developed a hybrid that was both simple and compatible to our onset detection analysis. We started by computing the timecourse of each of the three effects (fricative, vowel and coarticulation) for each subject. From these, we conducted a one-sample t-test at each 4 msec timeslice, asking whether the effect differed from 0 (a series of such t-tests is at the heart of both BDOTs, and Cluster-Based Permutation). We then corrected these t-tests using a family-wise error correction that uses the autocorrelation among T-statistics to identify an adjusted alpha value that is less conservative than a Bonferonni correction, and offers true family-wise error correction (unlike FDR approaches, Benjamini & Hochberg, 1985) (see, Oleson et al., 2017, for a derivation and Monte Carlo validation). This avoids the assumption of parametric curves made by BDOTs, and the assumptions about clustering in Cluster-based approaches.

Effects over time were smoothed with a 48 msec triangular window (this was not done for the onset analysis) and compared to 0. Figure 3A shows effect size (Cohen's D) of each of

the four effects over time. Table 2 summarizes the significant regions. All four effects had moderate to large effects (D>.5) at some time regions, though it was clear that the stop, fricative and vowel effects are much larger (particularly late in processing) than the effect of coarticulation. In terms of their rough timing, the stop effect appears to onset first roughly 80 msec after stimulus onset (adjusted for the oculomotor delay). Surprisingly, the fricative also shows a rapid onset at 116 msec (adjusted), however, we note that the size of the effect remains small until much later (around 300 ms). The effect of coarticulation began fairly early (284 ms), peaked, and disappeared by 448 ms. This disappearance likely reflects the fact that once the unambiguous vowel arrived, listeners stopped using the coarticulation in the frication. Finally, the vowel effect began late (260 msec, roughly immediately after vocoid onset). However, shortly after, it mirrored the fricative effect.

One source of uncertainty in this way of visualizing and analyzing the data is that since the stimuli were recordings of naturally produced utterances, the frications varied in length. This makes it difficult to precisely time lock effects to the vocoid (in order, for example, to determine if the anticipatory effect truly precedes the vowel). Thus, to obtain a clearer picture of the effects on trials with fricative groupings, we recomputed the effects over time, timelocked to the onset of the vocoid on each trial. These are shown in Figure 3B (the stop effect is not shown since the vocoid and stimulus onset are nearly coincident). Here again, the same pattern emerges, with an early small effect of frication that rapidly grows after vocoid onset mirroring that of the vowel. The effect of coarticulation was significant from 116 ms, when we factor in the oculomotor delay, this suggests it was present −84 msec prior to the vocoid (during the frication).

Comparison of Onsets

Our primary questions concerned differences among the timing of the effects. This preplanned analysis used the onset detection technique developed and validated by J. O. Miller et al. (1998) for comparing the onset of ERP components and used in most studies of the timecourse of cue integration (McMurray et al., 2008; Reinisch & Sjerps, 2013; Toscano & McMurray, 2012, 2015), including that of Galle et al. (in press), on which the present study builds.

Our ultimate goal was to estimate and compare the onset of each effect (e.g., the time the effect crossed a threshold). However, these effect-over-time curves were noisy for individual participants, leading to spurious variability in the estimated onsets. Thus, we jackknifed the data by averaging the timecourse of the effects across all participants, excluding one. Thresholds were then estimated on the basis of the jackknifed curves. This was repeated for each participant in turn to yield a new dataset the same size as the original (see J. O. Miller et al., 1998, for discussion and Monte Carlo simulations validating this technique for onset detection).

Since the question here concerned the timing of each effect, rather than its strength, the jackknifed effect sizes were then normalized by dividing the effect size at each time by the maximum of the effect (across the whole timecourse) for each participant. As a result, this normalized measure can be interpreted as the proportion of the maximal effect at any given moment. Finally, we estimated the onset of the effect by determining when the timecourse

crossed a fixed threshold (e.g., 30% of maximum). As in prior studies, this was conducted at multiple thresholds to ensure results are not idiosyncratic to a single one (McMurray et al., 2008; Reinisch & Sjerps, 2013). We selected the 20%, 30%, and 40% thresholds in advance of this study to match those used by (Galle et al., in press). Finally, the effect onsets were compared using a t-statistic which is modified to reflect the fact that after jackknifing estimates of the variance are much lower (R. G. Miller, 1974).

This approach overcomes three problems inherent in the repeated t-tests used in the prior analysis, in that 1) it computes an estimate of time for each participant (which can then be compared statistically); 2) it is less affected by the absolute size of the effect since effects are normalized; and 3) the estimate of effect onset is not biased by the variance (it is directly estimated from the mean effect for that subject), even as variance can make it more difficult to detect a difference.

Figure 4 shows the normalized time course of all four effects. Figure 3A shows time relative to the onset of the trial (100 msec prior to the onset of the fricative). Consistent with Galle et al. (in press), the fricative was identified late and nearly simultaneously with the vowel (though the prior analysis identified a small region that is above 0 earlier). The stops were also identified quite a bit earlier than either the fricative or the vowel. In contrast, coarticulation was used to identify the vowel much earlier than either the fricative or vowel, although its effect wanes later in the stimulus when the vocoid (containing more robust information about the vowel) is available.

Figure 4B shows the same data with time relative to vocoid onset. Again, we see relatively late onsets for both the fricative and vowel effects. But, note here that since 0 ms is the onset of the vocoid, the vowel effect and the bulk of the fricative effect arises roughly 200 ms *after* vocoid onset. Given the fact that it takes roughly 200 ms to plan and launch an eye-movement (Viviani, 1990), this suggests the fixations that ultimately underlie these effects were roughly planned at vocoid onset. This confirms late integration of cues to fricative identification. In contrast, the coarticulatory effect departs from 0 at around 0 ms; given the 200 ms planning delay, this would suggest that it was planned prior to vocoid onset, during the frication. This suggests immediate utilization of coarticulation. This visualization also makes it clear that the small region of significance in the fricative effect (identified in the preceding analysis) that appears early (during the frication) is only a small portion of how much the fricative will eventually be used (after the vocoid when the curve takes a sharp bend).

To analyze these statistically, the onset of each effect (at each pre-determined threshold) was compared using a series of paired-t tests (adjusted for the use of the jackknife procedure) which compared the estimated onset of all three effects. We conducted two types of comparisons. First, to replicate Galle et al. (in press), we compared the effect onsets of the stop and fricative trials in absolute time. Second, to extend it, we compared the onset of the fricative effect to that of the vowel effect in relative time to determine if fricative identification is timelocked to vowel identification or just to the vocoid. Finally, we analyzed our planned contrasts for the coarticulation effects by comparing the onsets of the fricative, vowel, and coarticulatory effects, when these effects were computed relative to the vocoid

onset. As our primary statistical tests were intended to be conducted in vocoid-relative time, we avoided duplicating these results in the initial absolute time analysis to minimize the number of tests.

Table 3 shows comparisons made in absolute time. On average, the fricative effect began roughly 150 ms later than the stop voicing effects. This replicates the delayed onset of the fricative effect found by (Galle et al., in press). One might argue that this is an unfair comparison since stops are shorter than fricatives. However, gating studies suggest that fricatives can be identified at very high accuracy rates with extremely short gates (e.g., 50 ms) (Galle et al., in press; Smits, Warner, McQueen, & Cutler, 2003; Warner, McQueen, & Cutler, 2014). There is ample information available very early to support fricative discrimination, yet listeners appear to wait to use it to access the lexicon.

Table 4 shows results of our primary comparisons which treated time relative to vocoid onset. First, we compared the fricative to the vowel effect, to extend Galle et al. (in press). Both effects began about 300 ms after the onset of the vocoid; given the 200 ms oculomotor delay, this is shortly after the vocoid's onset. At the 30%, and 40% thresholds there were no significant differences between the onset of the fricative and vowel effects. While the fricative effect was significantly earlier than the vowel effect at the 20% threshold, the difference was numerically very small (31 ms), relative to the actual time difference between these events in the signal (264 ms), suggesting this early effect does not reflect truly immediate processing of the fricative. This extends the Galle et al. (in press) findings by suggesting that fricative identification is occurring largely concurrently with vowel identification (though there may be small effects that precede it).

Finally, we addressed the primary question: the onset of the coarticulation effect. The onset of the coarticulation effect was about 30 ms after the physical onset of the vocoid. Given the assumed 200 ms oculomotor delay, this is ~170 ms before vocoid onset. As the fricatives averaged 264 ms in length, this is as early as could be expected. This is supported by phonetic analyses (Supplementary Analysis S2) that suggest that there is reliable coarticulatory information present in the first 50 ms of both the /s/ and /ʃ/ tokens. Critically, the effect of coarticulation began significantly before both the vowel and fricative effects at all three thresholds, by a substantial period of time (~300 ms).

General Discussion

Our first finding replicates Galle et al. (in press): lexical activation did not reflect the distinction between the fricatives for quite some time. Listeners showed significantly later utilization of frication cues at word onset than stop voicing cues (Figure 4A, Table 3). We should note that exploratory analyses identified a small, but significant effect of frication early (consistent with secondary analyses in Galle et al., in press, Supplement S2). However, this effect was small (less than 10% of what the total effect size would be) and did not reach the 20% threshold identified planned, conventional analysis. This is now the sixth experiment showing this delayed effect. Moreover, we extend that study Galle et al. (in press) by showing that the timing of the frication effect was not different from that of the vowel (Figure 4B Table 4). Both occurred roughly 200 ms after the onset of the vocoid,

suggesting a tight coupling of the consonants and vowel decisions. These two findings support a late integration account for fricative identification and we briefly discuss that in the next section.

The more important and novel finding from this study is that listeners use coarticulatory information in the frication to anticipate vowels before much of the information in the frication is used to identify the fricative (and before the information in the vocoid is used to identify the vowel). This finding should be moderated by the fact that, as expected, the anticipatory effect is small in absolute terms (Figure 3). Nonetheless, the presence of a significant difference in onset suggests an immediate utilization strategy for anticipatory coarticulation, even as listeners use a largely late integration strategy for fricative identification. Thus, listeners begin to use information in the frication to make decisions about the subsequent vowel before any vowel is present in the input, and critically, *before they do so for the fricative*.

In the remainder of this discussion we briefly discuss the late integration of fricatives before discussing the early utilization of coarticulation. Finally, we turn to broader—and more speculative—thoughts about how to conceptualize time in speech perception more generally.

Late Integration of Fricatives

Our first two findings replicate and extend Galle et al. (in press), showing late integration of sibilant fricatives—listeners appear to wait to make the bulk of the decision about a fricative until the onset of the vocoid, and this decisions are closely timelocked with decisions about the vowel (on the basis of the vocoid). We note that this is not an argument from a null effect. Frication was used significantly later than the onsets of stop consonants. Further, Galle et al. (in press) experimentally manipulated the length of the frication, and showed that lengthening the frication created a concomitant delay in the fricative decision. As we described in the introduction, late integration has been extensively validated by that study, ruling out a large number of experimental factors. As this is not the primary finding here, we leave a more thorough discussion to that paper. However, two points are worth noting.

The first issue is whether this finding of late integration of frication cues (and by extension the early anticipation) is limited to voiceless sibilants. At the moment this is unclear. We suspect that voiceless non-sibilants (like /f/ and /θ/) will show similar effects, but for unsurprising reasons: for non-sibilants (unlike sibilants), the frication alone is rarely sufficient to identify the fricative (Jongman et al., 2000; McMurray & Jongman, 2011). One might also observe this effect with long aspirated stop consonants (e.g., in Navaho), or with voiceless vowels (in Japanese)—in both cases the vocoid is either late or absent. Voiced fricatives pose an interesting question, as it is not clear whether it is the vocoid that is needed to release the contents of the buffer, or the onset of any voicing. This would be difficult to test as /ʒ/ does not appear word initially in English. However, the broader point of our work with fricatives is not whether people use late integration more generally for everyday speech perception (they likely do not for most sounds). Rather, the presence of late integration for any phonemes suggests the need for a much more complex architecture for speech perception than previously thought.

The second issue is whether this somehow an artifact of the task. How would the interpretation of these results change if fixations reflect response planning, not lexical access? To some extent, we would argue that fixations do reflect response planning, as we are using a "goal based" version of the VWP (Salverda, Brown, & Tanenhaus, 2011) in which people must make an overt response on every trial, and eye-movements must reflect this behavior (not a simple match of whatever is active in the lexicon to whatever is available in the visual scene). However, such planning must be filtered through the lexical/semantic system since responses reflect meaningful pictures. In this case, it is not clear why the lexical/semantic system would withhold preliminary activation states for some words (fricative-initial), but not others (stop-initial) as the meanings of these words don't differ systematically. Rather, it seems more reasonable to locate this buffering in the perceptual processes that precede lexical access, given the strong acoustic/phonetic differences. Moreover, the two studies that purport to show immediate activation of fricatives (Kingston et al., 2016; Mitterer & Reinisch, 2013), as well as most gating studies (Galle et al., in press; Wagner, Ernestus, & Cutler, 2006) bypass the lexical system with orthographic responses. This suggests a locus in the mapping between the signal and the lexicon. Finally, we note that Galle et al. (in press) ruled out several task and stimulus confounds, and that their study as well as the present used an identical task to a range of studies showing immediate integration (McMurray et al., 2008; Toscano & McMurray, 2012, 2015). Thus, this is unlikely to derive from global task properties.

While there are clearly further questions remaining, these results along with those of Galle et al. (in press) clearly support the hypothesis that voiceless sibilant fricatives are integrated late.

Anticipation and Cue Integration

The primary novel finding of this study is the relative early utilization of anticipatory coarticulation in the frication. The early utilization of coarticulation is not in and of itself surprising—it has previously been shown for assimilation in stops and nasals (Gow, 2001) and for coarticulation in preceding vowels (Salverda et al., 2014). However, given the late utilization of frication, it was unexpected in this particular context: people appear to be able to use coarticulatory information in the frication to anticipate subsequent vowels *before* they identify the current fricative. This is supported by prior gating work (Nittrouer & Whalen, 1989) that has identified situations in which listeners may be able to identify the fricative but not anticipate the vowel, or vice versa. Thus, these decisions are clearly not dependent upon each other (though they may benefit each other: McMurray & Jongman, 2015).

Galle et al. (in press) discuss several possible reasons for why retroactive cue integration in sibilants are characterized by late integration. Our results on anticipation may clarify this. The most prominent theory they discuss is that late integration derives from the acoustic properties of these speech sounds, the fact that information must be integrated across disparate frequency bands (consistent with a largely auditory account of speech perception Diehl et al., 2004). The present study rules this out. The coarticulatory information in the signal is largely based in these high frequency ranges (see Supplement S3) and is intended to predict the vowel (largely in lower frequencies). Yet, the anticipatory effect suggests same

portion of the signal can be used immediately to identify the vowel, even as it is delayed for the fricative. This suggests that late integration is a product of what contrast is operative (the fricative vs. the vowel), not the acoustic input which listeners use to make the contrast.

Similarly, Galle et al. (in press) also raise the possibility that late integration derives from the fact that fricatives are contextually dependent on the vocoid: knowing that the vowel was rounded or unrounded or that the talker was male or female will help identify the fricative (Apfelbaum et al., 2014; McMurray & Jongman, 2011; Strand, 1999); consequently listeners may wait for the vocoid to identify the fricatives. This too is ruled out by the present data. Coarticulatory cues in the frication are not that useful in predicting the vowel (c.f. Figure 3A), and listeners can get a lot more out of them if they have access to contextual information like talker (McMurray & Jongman, 2015). Yet listeners use them immediately. In contrast, gating studies suggest that even very short segments of the frication are sufficient for highly accurate fricative identification (Galle et al., in press; Smits et al., 2003; Warner et al., 2014), and yet they are used late. Therefore, the relative contextual dependence of various cues does not offer a clear explanation for this phenomenon.

In contrast, both the late integration of fricative cues, and the early utilization of anticipatory coarticulation is consistent with an account in which the vowel is the organizing unit for each syllable (Diehl et al., 1987). This idea appears in older account like the P-Center account (Marcus, 1981), and the present study was not meant as a definitive test of that hypothesis. But of the hypotheses outlined in Galle et al. (in press), it seem the only one that can accounts for the present results. Under this view, vowels must be identified before the rest of the syllable can be constructed. If so, fricatives may not be identified until late—when the vocoid arrives. This can thus account for the late utilization of frication.

So how can this account for immediate utilization in other types of sounds? Stops are both shorter and contain substantial information about the vowel. For example, studies on silent center vowels show that listeners can categorize vowels equivalently well with stop consonant formant transitions alone as they can with the complete vowel (Jenkins et al., 1983; Parker & Diehl, 1984; Strange, Jenkins, & Johnson, 1983). Given that vowel information is thus present from the earliest portions of the words, the vowel may be accessed immediately, allowing listeners to then immediately integrate cues for other phonemes. It is unclear what would happen with longer consonants (nasals and liquids), though this might depend on whether the vowel could be identified based on early coarticulatory information.

Can a vowel-centered effect explain the early utilization of anticipatory coarticulation? Here, the need to identify the vowel could lead listeners to use coarticulation to anticipate the vowel immediately (that’s the first thing that needs to happen), before any other segments can be identified. Thus, listeners may use coarticulation early to get a head start on vowel identification; but only integrate the fricative later when the vowel comes on line.

Vowel centered accounts are old, but there are studies that provide converging evidence (to our own) for the centrality of the vowel. For example, Diehl et al. (1987) showed that listeners RTs to a judgement of word-initial consonants were correlated with the inherent

length of the vowel, suggesting some contingency between consonant and vowel identification. And emerging work using direct recordings from the cortices of epilepsy patients suggests targeted cortical sites within speech processing areas respond to rapid changes in the envelope (e.g., as would be found at the juncture between a fricative and a vocoid), suggesting a unique role for voicing onset (Oganian & Chang, 2018). However, unlike older vowel-centered accounts, our studies suggest that it is not the center or the end of the vowel that is the anchor for perceptual analysis; rather listeners initiate perceptual analysis whenever there is sufficient information available to identify the vowel.

Beyond this somewhat speculative vowel-centered approach there are broader implications for theories of speech. Up to this point, work on retrospective cue-integration and anticipation suggests a similar type of process for each class of effects: immediate utilization. People use information to make partial inferences as soon as it arrives, whether they are using this to make inferences about a current segment or about future ones. However, taken with the results of Galle et al. (in press), this story may be not be correct in all cases. The present study suggests listeners can use a different strategy for retrospective cue integration than for prospective or anticipatory processing, even in the context of the same chunk of input (the frication). This suggests the need for thinking about alternative conceptualizations of time.

Temporal Processing in Speech

Our study was narrowly focused on when various sorts of information is used in speech perception. However, in the spirit of a special issue dedicated to Randy Diehl, it is worth briefly surfacing to think more broadly. Our study was not designed to test broader theoretical accounts of time, so this discussion is speculative (or perhaps should be read as a critical commentary). However, the implications of this study (and related ones) are important for theories of speech perception.

Most accounts of temporal processing in speech derive from work on spoken word recognition, which has been framed around the problem of integrating information over time to recognize words (Dahan & Magnuson, 2006; Weber & Scharenborg, 2012). Early models assumed strict left to right processing, roughly timelocked to the input (Marslen-Wilson, 1987). As each phoneme arrives, words that contain this phoneme in the right position are activated (or deactivated if they do not). The more recent consensus has replaced this with a more flexible form of competition in which incoming speech activates or deactivates words in a partial or graded way (Hannagan, Magnuson, & Grainger, 2013; McClelland & Elman, 1986), with a form of dynamic competition unfolding over time. Competition models are not as closely timelocked to the input as older conceptualizations. Processes like inhibition, feedback between levels of processing, or self-sustaining activity within a level (or its inverse, decay) can lead words to be suppressed earlier, or maintained longer than would be predicted by the input alone and can lead to activation of words that are not fully timelocked to the input (e.g., rhymes).

However, at the broadest level, the dynamics of these models are strongly dictated by the dynamics of the input (Figure 5A). Words are represented in terms of the order of their constituents, and the unfolding input constrains activation to be largely consistent with left-

to-right processing. Moreover, such models all assume continuous cascades among levels of processing—preliminary states of auditory processing immediately affect phonemic processing, and preliminary states of phonemic processing immediately affect lexical processing.

As a result of such commitments, competition models are consistent with all the prior work demonstrating immediate utilization of acoustic cues (Kingston et al., 2016; McMurray et al., 2008; Mitterer & Reinisch, 2013; Reinisch & Sjerps, 2013; Toscano & McMurray, 2012, 2015). In contrast, the present results challenge this dominant view in two ways. First, our finding of late integration of frication cues prompts the need for some form of memory (to hold onto the frication cues while the listener waits for the vocoid) that is encapsulated from higher level processing (it does not continuously cascade to affect lexical access). Second, and perhaps more importantly, time does not seem to run linearly here (cf., Vonnegut, 1969). Listeners appear to make decisions about fricatives and vowels in the opposite order from which they arrived. While processes like inhibition and decay in competition models can partially decouple lexical activation from the dynamics of the input, it is not clear that any such parameter settings could decouple them to this extent.

This is perhaps the most dramatic documentation that the dynamics of the input do not fully determine the dynamics of the system. However, a number of broader finding converge on the need to consider time differently. It is important to point out that this was predicted by the results of Diehl et al. (1987); while they did not use an online measure such as used here, their data supporting a vowel-centered view of speech perception suggests processing may need to skip about in time. This is also hinted at by models of speech perception in which decisions about one phoneme may need to be predicated by earlier or later decisions (McMurray & Jongman, 2011; Smits, 2001), for example, when a fricative decision must be predicated by the later identity of the vowel.

In addition, several more recent findings challenge a general left-to-right (or incremental processing) framing in perhaps more direct ways. First, it should be noted that incremental or immediate word recognition develops very slowly, from infancy through late adolescence (Fernald, Perfors, & Marchman, 2006; Law, Mahr, Schneeberg, & Edwards, 2017; Rigler et al., 2015). While these studies all support incremental processing, the dramatic changes in the speed of incremental processing suggests that the internal dynamic of word recognition are not driven solely by the input; rather there are some internal dynamics that change over development.

Second, pre-linguistically deafened people who use cochlear implants, and normal hearing adults under highly degraded conditions, do not process spoken words immediately (McMurray et al., 2017). Rather, in both cases, listeners appear to wait several hundred milliseconds to initiate lexical access, and as a result they show *reduced* competition from onset competitors (e.g., when hearing *sandal,* they show less competition from *sandwich).* That is, by waiting, they accumulate enough information to partially rule out competitors before initiating lexical access (minimizing temporary ambiguity). This too suggests a language system whose dynamics are somewhat decoupled from the dynamics of the input, and may require some type of auditory buffer (as in Galle et al., in press).

Finally, Toscano, Anderson & McMurray (2013) showed that normal listeners activate phonemic anadromes like *cat* after hearing *tack.* This is not predicted by most current models of spoken word recognition (and TRACE simulations failed to find a version of the model which could account for this results). This is because most models propose that lexical representations (either explicitly or implicitly) code the position (slot) in a word in which each phoneme should appear and use this to constrain lexical access. By such a slot-based account, a /k/ word initially (as in *cat*) should activate other words with a /k/ in initial position (*can, cap*), but cannot serve to activate words with the /k/ in final position (*tack).* Indeed, by these slot-based accounts, *cat* is as distinct from *tack* as it is from *pad* (where both consonants mismatch). Thus, Toscano's results suggest that phonemic information within a word may be only coarsely coded by time.

All three of these findings could potentially be handled by existing interactive activation or competition models with appropriate modifications (Hannagan et al., 2013; McClelland & Elman, 1986) by tuning processes like inhibition or decay (among words or phonemes), that can control the internal dynamics. At appropriate parameter settings, these may be able to handle the gains over development, and the delays seen with severely degraded speech (cf., McMurray, Samelson, Lee, & Tomblin, 2010, for simulations with many parameter settings). Further, alternative patterns of connectivity (the lexical "representation") could support a less slot-like approach to temporal order (Hannagan et al., 2013) to account for Toscano, Anderson, and McMurray (2013).

However, the present results may push these findings even farther, showing a more dramatic departure from typical left-to-right processing: listeners have made a partial decision about the second phoneme (the vowel) before they have made a decision on the first one (the fricative). This suggests that simply tweaking the dynamics of lexical competition models is insufficient and that a different approach to time may be able to account for all of these results more gracefully.

To be clear, we are not proposing that the vowel centered approach that may account for the results of the present study underlies these phenomena. Each of these phenomena may require a unique approach to time. However, looking more broadly, these studies suggest that the dynamics of the cognitive system may only be loosely coupled to the dynamics of the input (Figure 5B), though for various reasons. Under this view, word recognition (or language processing) doesn't "start" when the first input arrives; rather, language processing is ongoing by its own internal dynamics, and these can be pushed around as new input arrives. These ideas are similar to ideas proposed by (Elman, 2009), though we note that his specific model could not likely account for our effects as it has no way to buffer acoustic processing from the rest of language processing. Such an approach could be advantageous for integrating long term expectations (about a talker or dialect, or from semantic or discourse context with ongoing perceptual processing), as such expectations may have a slower timescale than the real-time arrival of the input and may not change as rapidly. However, even this may be too simple: the fact that early fricative decisions must be held in a buffer until the vowel arrives suggests that there may be multiple independent "tiers" of analysis (perceptual, lexical, sentential), each with their own internal dynamics, and each only loosely coupled to the dynamics of the input.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Jamie Klein-Packard and the members of the MACLab for assistance with data collection; Marcus Galle for critical ideas leading to this work; Mathias Sjerps for suggesting the P-Center hypothesis, and Kilgore Trout for assistance with the title of this project. We would also like to thank Randy Diehl whose lifetime of scholarship demonstrated a commitment to questioning assumptions, thinking deeply about the role of basic auditory processing in speech, and exploring odd phenomena to lasting theoretical gain. This research was supported by DC 0008089 awarded to BM.

Appendix:: Complete word list and their assignment to condition.

Note that this shows assignments of items to conditions only when the “target” is the auditory stimulus. When the competitor is the auditory stimulus, the target then becomes the competitor. Similarly, when either of the unrelated items are the auditory stimulus the trial was a filler trial and was not analyzed.

	Grouping					
Set	**Contrast**	**Context**	**Target**	**Competitor**	**Unrelated**	**Unrelated**
1	Fricative	*Unrounded*	*seed*	*sheep*	*Bean*	*pail*
		Rounded	*soup*	*shoot*	*Bite*	*pit*
	Vowel	/ʃ/	*sheep*	*shoot*	*Pail*	*pit*
		/s/	*seed*	*soup*	*Bean*	*bite*
2	Fricative	*Unrounded*	*seat*	*sheet*	*Bug*	*pill*
		Rounded	*suit*	*shoes*	*Back*	*pat*
	Vowel	/ʃ/	*sheet*	*shoes*	*Pill*	*pat*
		/s/	*seat*	*suit*	*Bug*	*back*
3	Fricative	*Unrounded*	*cent*	*shell*	*Bath*	*pal*
		Rounded	*sword*	*shore*	*Beer*	*pine*
	Vowel	/ʃ/	*shell*	*shore*	*Pal*	*pine*
		/s/	*cent*	*sword*	*Bath*	*beer*
4	Fricative	*Unrounded*	*sip*	*shin*	*Butt*	*path*
		Rounded	*soot*	*shook*	*bone*	*pick*
	Vowel	/ʃ/	*shin*	*shook*	*path*	*pick*
		/s/	*sip*	*soot*	*butt*	*bone*
5	Fricative	*Unrounded*	*cell*	*shed*	*bell*	*peak*
		Rounded	*soap*	*show*	*bear*	*peg*
	Vowel	/ʃ/	*shed*	*show*	*peak*	*peg*
		/s/	*cell*	*soap*	*bell*	*bear*
6	Fricative	*Unrounded*	*self*	*chef*	*bet*	*pack*
		Rounded	*sore*	*shorts*	*bum*	*pig*

	Grouping					
Set	**Contrast**	**Context**	**Target**	**Competitor**	**Unrelated**	**Unrelated**
	Vowel	/ʃ/	*chef*	*shorts*	*pack*	*pig*
		/s/	*self*	*Sore*	*bet*	*bum*

References.

Altmann GTM, & Kamide Y (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73(3), 247–264. doi:10.1016/s0010-0277(99)00059-1 [PubMed: 10585516]

Apfelbaum KS, Blumstein SE, & McMurray B (2011). Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. Psychonomic Bulletin and Review, 18(1), 141–149. [PubMed: 21327343]

Apfelbaum KS, Bullock-Rest N, Rhone A, Jongman A, & McMurray B (2014). Contingent categorization in speech perception. Language, Cognition and Neuroscience, 29(9), 1070–1082.

Beddor PS, Harnsberger JD, & Lindemann S (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. Journal of Phonetics, 30(4), 591–627.

Benjamini Y, & Hochberg Y (1985). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 85, 289–300.

Boersma P, & Weenink D (2009). Praat: doing phonetics by computer (Version Version 5.1.05) Retrieved from http://www.praat.org/

Bregman A (1990). Auditory Scene Analysis: The Perceptual Organization of Sound Cambridge, MA: The MIT Press.

Clayards M, Tanenhaus MK, Aslin RN, & Jacobs RA (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804–809. [PubMed: 18582855]

Cole JS, Linebaugh G, Munson C, & McMurray B (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. Journal of Phonetics, 38(2), 167–184. [PubMed: 21173864]

Dahan D, & Magnuson JS (2006). Spoken-word recognition. In Traxler MJ & Gernsbacher MA (Eds.), Handbook of Psycholinguistics (pp. 249–283). Amsterdam: Academic Press.

Daniloff R, & Moll K (1968). Coarticulation of Lip Rounding. Journal of Speech, Language, and Hearing Research, 11(4), 707–721. doi:10.1044/jshr.1104.707

Diehl RL, Kluender KR, Foss DJ, Parker EM, & Gernsbacher MA (1987). Vowels as islands of reliability. Journal of Memory and Language, 26(5), 564–573. doi: 10.1016/0749-596X(87)90143-4 [PubMed: 25505818]

Diehl RL, Lotto AJ, & Holt LL (2004). Speech Perception. Annual Review of Psychology, 55(1), 149–179. doi:10.1146/annurev.psych.55.090902.142028

Diehl RL, & Walsh MA (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. Journal of the Acoustical Society of America, 85, 2154–2164. [PubMed: 2732389]

Elman JL (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. Cognitive Science, 33, 547–582. [PubMed: 19662108]

Fernald A, Perfors A, & Marchman VA (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology, 42(1), 98–116. [PubMed: 16420121]

Forrest K, Weismer G, Milenkovic P, & Dougall RN (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. Journal of the Acoustical Society of America, 84, 115–124. [PubMed: 3411039]

Frazier L (1987). Sentence processing: A tutorial review

Galle ME, Klein-Packard J, Schreiber K, & McMurray B (in press). What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. Cognitive Science

Gaskell MG, & Marslen-Wilson WD (1997). Integrating form and meaning: a distributed model of speech perception. Language and Cognitive Processes, 12(5/6), 613–656.

Gow DW (2001). Assimilation and Anticipation in continuous spoken word recognition. Journal of Memory and Language, 45, 133–139.

Gow DW (2003). Feature parsing: Feature cue mapping in spoken word recognition. Perception & Psychophysics, 65(4), 575–590. [PubMed: 12812280]

Gow DW, & McMurray B (2007). Word recognition and phonology: The case of English coronal place assimilation. In Cole JS & Hualdo J (Eds.), Papers in Laboratory Phonology 9 (pp. 173–200). New York, NY: Mouton de Gruyter.

Hannagan T, Magnuson J, & Grainger J (2013). Spoken word recognition without a TRACE. Frontiers in Psychology, 4(563). doi:10.3389/fpsyg.2013.00563

Hawkins S (2003). Roles and representations of systematic fine phonetic detail in speech understanding. Journal of Phonetics, 31, 373–405.

Hoequist CE (1983). The Perceptual Center and Rhythm Categories. Language and Speech, 26(4), 367–376. [PubMed: 6677830]

Holt LL, & Lotto AJ (2008). Speech Perception Within an Auditory Cognitive Science Framework. Current Directions in Psychological Science, 17(1), 42–46. doi:10.1111/j.1467-8721.2008.00545.x [PubMed: 19060961]

Jenkins JJ, Strange W, & Edman TR (1983). Identification of vowels in “vowelless” syllables. Perception & Psychophysics, 34(5), 441–450. doi:10.3758/bf03203059 [PubMed: 6657448]

Jenkins JJ, Strange W, & Miranda S (1994). Vowel identification in mixed- speaker silent- center syllables. The Journal of the Acoustical Society of America, 95(2), 1030–1043. doi: 10.1121/1.410014 [PubMed: 8132897]

Jongman A, Wayland R, & Wong S (2000). Acoustic characteristics of English fricatives. Journal of the Acoustical Society of America, 106, 1252–1263.

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. [PubMed: 27736119]

Law F, Mahr T, Schneeberg A, & Edwards J (2017). Vocabulary size and auditory word recognition in preschool children. Applied Psycholinguist, 38(1), 89–125.

Levy R, Bicknell K, Slattery T, & Rayner K (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. Proceedings of the National Academy of Sciences, 106(50), 21086–21090.

Luck SJ, & Vogel EK (1997). The capacity of visual working memory for features and conjunctions. Nature, 390(6657), 279. [PubMed: 9384378]

MacDonald MC, Pearlmutter NJ, & Seidenberg MS (1994). Lexical nature of syntactic ambiguity resolution. Psychological Review, 101, 676–703. [PubMed: 7984711]

Magen HS (1997). The extent of vowel-to-vowel coarticulation in English. Journal of Phonetics, 25(2), 187–205.

Mann VA, & Repp B (1980). Influence of the vocalic context on the ? - s distinction. Perception & Psychophysics, 28(3), 213–228. [PubMed: 7432999]

Marcus SM (1981). Acoustic determinants of perceptual center (P-center) location. Perception & Psychophysics, 30(3), 247–256. [PubMed: 7322800]

Maris E, & Oostenveld R (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164(1), 177–190. doi:10.1016/j.jneumeth.2007.03.024 [PubMed: 17517438]

Marslen-Wilson WD (1987). Functional parallelism in spoken word recognition. Cognition, 25(1–2), 71–102. [PubMed: 3581730]

McClelland JL, & Elman JL (1986). The TRACE model of speech perception. Cognitive Psychology, 18(1), 1–86. [PubMed: 3753912]

McMurray B, Clayards M, Tanenhaus MK, & Aslin RN (2008). Tracking the time course of phonetic cue integration during spoken word recognition. Psychonomic Bulletin and Review, 15(6), 1064–1071. [PubMed: 19001568]

McMurray B, Farris-Trimble A, & Rigler H (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. Cognition, 169, 147–164. [PubMed: 28917133]

McMurray B, & Jongman A (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review, 118(2), 219–246. [PubMed: 21417542]

McMurray B, & Jongman A (2015). What comes after [f]? Prediction in speech is a product of expectation and signal. Psychological Science, 27(1), 43–52. [PubMed: 26581947]

McMurray B, Samelson VS, Lee SH, & Tomblin JB (2010). Individual differences in online spoken word recognition: Implications for SLI. Cognitive Psychology, 60(1), 1–39. [PubMed: 19836014]

McMurray B, Tanenhaus MK, & Aslin RN (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. Journal of Memory and Language, 60(1), 65–91. [PubMed: 20046217]

Miller JL, & Dexter ER (1988). Effects of speaking rate and lexical status on phonetic perception. Journal of Experimental Psychology: Human Perception and Performance, 14(3), 369. [PubMed: 2971767]

Miller JL, & Volaitis LE (1989). Effect of speaking rate on the perceptual structure of a phonetic category. Perception & Psychophysics, 46(6), 505–512. [PubMed: 2587179]

Miller JO, Patterson T, & Ulrich R (1998). Jackknife-based method for measuring LRP onset latency differences. Psychophysiology, 35, 99–115. [PubMed: 9499711]

Miller RG (1974). The jackknife-a review. Biometrika, 61(1), 1–15. doi:10.1093/biomet/61.1.1

Mitterer H, & Reinisch E (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. Journal of Memory and Language, 69(4), 527–545.

Nearey TM, & Rochet BL (1994). Effects of Place of Articulation and Vowel Context on VOT Production and Perception for French and English Stops. Journal of the International Phonetic Association, 24(1), 1–18.

Nittrouer S, & Whalen DH (1989). The perceptual effects of child–adult differences in fricative- vowel coarticulation. The Journal of the Acoustical Society of America, 86(4), 1266–1276. doi: 10.1121/1.398741 [PubMed: 2808902]

Oganian Y, & Chang EF (2018). A speech envelope landmark for syllable encoding in human superior temporal gyrus. bioRxiv, 10.1101/388280

Ohde RN (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. Journal of the Acoustical Society of America, 75(1), 224–230. [PubMed: 6699284]

Ohman SEG (1966). Coarticulation in VCV utterances: Spectrographic measurements. Journal of the Acoustical Society of America, 39, 151–168. [PubMed: 5904529]

Oleson JJ, Cavanaugh JE, McMurray B, & Brown G (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. Statistical Methods in Medical Research, 26(6), 2708–2725. doi:10.1177/0962280215607411 [PubMed: 26400088]

Parker EM, & Diehl RL (1984). Identifying vowels in CVC syllables: Effects of inserting silence and noise. Perception & Psychophysics, 36(4), 369–380. doi:10.3758/bf03202791 [PubMed: 6522234]

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

Rigler H, Farris-Trimble A, Greiner L, Walker J, Tomblin JB, & McMurray B (2015). The slow developmental timecourse of real-time spoken word recognition. Developmental Psychology, 51(12), 1690–1703. [PubMed: 26479544]

Salverda AP, Brown M, & Tanenhaus MK (2011). A goal-based perspective on eye movements in visual world studies. Acta Psychologica, 137(2), 172–180. [PubMed: 21067708]

Salverda AP, Kleinschmidt D, & Tanenhaus MK (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. Journal of Memory and Language, 71(1), 145–163. [PubMed: 24511179]

Seedorff M, Oleson JJ, & McMurray B (in press). Detecting when timeseries differ: Using the Bootstrapped Differences of Timeseries (BDOTS) to analyze Visual World Paradigm data (and more). Journal of Memory and Language

Sereno JA, Baum SR, Marean GC, & Lieberman P (1987). Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children. The Journal of the Acoustical Society of America, 81, 512. [PubMed: 3558969]

Smits R (2001). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. Perception & Psychophysics, 63, 1109–1139. [PubMed: 11766939]

Smits R, Warner N, McQueen JM, & Cutler A (2003). Unfolding of phonetic information over time: a database of Dutch diphone perception. Journal of the Acoustical Society of America, 113(1), 563–574. [PubMed: 12558292]

Strand E (1999). Uncovering the Role of Gender Stereotypes in Speech Perception. Journal of Language and Social Psychology, 18, 86–100.

Strange W, Jenkins JJ, & Johnson TL (1983). Dynamic specification of coarticulated vowels. The Journal of the Acoustical Society of America, 74(3), 695–705. 10.1121/1.389855 [PubMed: 6630725]

Summerfield Q (1981). Articulatory rate and perceptual constancy in phonetic perception. Journal of the Acoustical Society of America, 7(5), 1074–1095.

Summerfield Q, & Haggard M (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 62(2), 435–448. [PubMed: 886081]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, & Sedivy JC (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632–1634. [PubMed: 7777863]

Toscano JC, Anderson ND, & McMurray B (2013). Reconsidering the role of temporal order in spoken word recognition. Psychonomic Bulletin & Review, 20, 1–7. [PubMed: 23090749]

Toscano JC, & McMurray B (2012). Online integration of acoustic cues to voicing: Natural vs. synthetic speech. Attention, Perception & Psychophysics, 74(6), 1284–1301.

Toscano JC, & 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, 529–543.

Viviani P (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In Kowler E (Ed.), Eye Movements and Their Role in Visual and Cognitive Processes. Reviews of Oculomotor Research V4 (pp. 353–383). Amsterdam: Elsevier.

Vonnegut K (1969). Slaughterhouse-Five, or The Children's Crusade: A Duty-Dance with Death New York, NY: Delacorte Publishing.

Wagner A, Ernestus M, & Cutler A (2006). Formant transitions in fricative identification: The role of native fricative inventory. The Journal of the Acoustical Society of America, 120(4), 2267–2277. doi:10.1121/1.2335422 [PubMed: 17069322]

Warner N, McQueen JM, & Cutler A (2014). Tracking perception of the sounds of English. Journal of the Acoustical Society of America, 135(5), 2995–3006. doi:10.1121/1.4870486 [PubMed: 24815279]

Warren P, & Marslen-Wilson W (1987). Continuous uptake of acoustic cues in spoken word recognition. Perception & Psychophysics, 41(3), 262–275. doi:10.3758/bf03208224 [PubMed: 3575084]

Weber A, & Scharenborg O (2012). Models of spoken-word recognition. Wiley Interdisciplinary Reviews: Cognitive Science, 3(3), 387–401. doi:10.1002/wcs.1178 [PubMed: 26301470]

Yeni–Komshian GH, & Soli SD (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative‐ vowel coarticulation. The Journal of the Acoustical Society of America, 70, 966. [PubMed: 7288043]

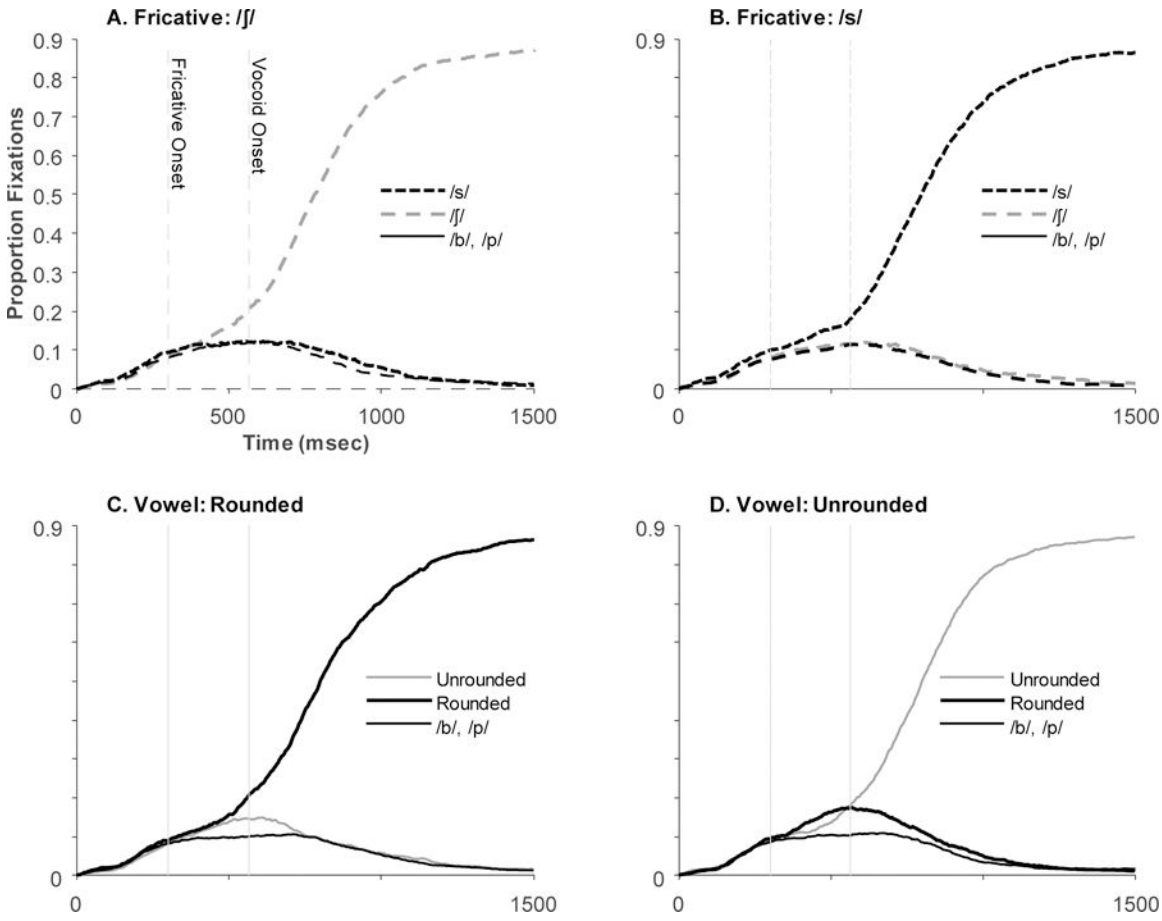

Figure 1.

Proportion fixations to each object as a function of time. Note that in all plots looks to the fillers (/b/, /p/) represent the average of the two objects. Time is relative to trial onset and not adjusted for the oculomotor delay. Thus, 300 msec (indicated by vertical line) is the first point at which a stimulus driven response could occur. Vertical lines represent fricative onset and the mean vocoid onset after adjusting for the oculomotor delay. A) For fricative groupings when the stimulus was /ʃ/-initial, averaged across both vowels and both coarticulatory conditions; B) For fricative groupings when the stimulus was /s/-initial; C) For vowel groupings when the stimulus contained a rounded vowel (averaged across both fricatives and coarticulation conditions); D) For vowel groupings when the stimulus contained an unrounded vowel.

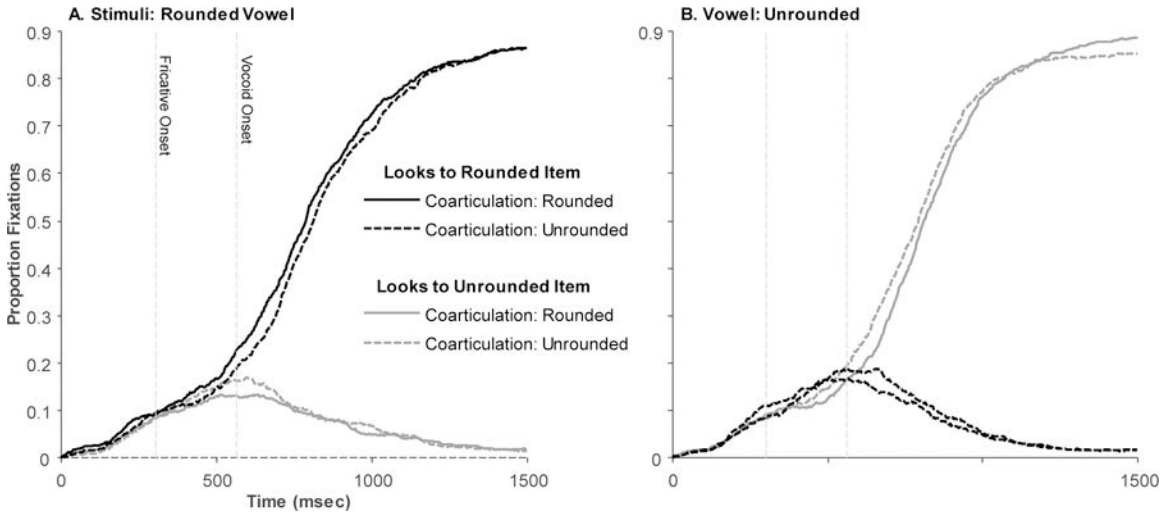

Figure 2.

Effect of anticipatory coarticulation on vowel groupings. Shown are looks to the unrounded (gray lines) vs. rounded (black lines) object as a function of coarticulation (dashed vs. solid lines). Time is relative to trial onset and not adjusted for the oculomotor delay. Vertical lines represent fricative onset and the mean vocoid onset (after adjusting for the 200 msec oculomotor delay). A) When the stimulus included a rounded vowel. B) When the stimulus included an unrounded vowel.

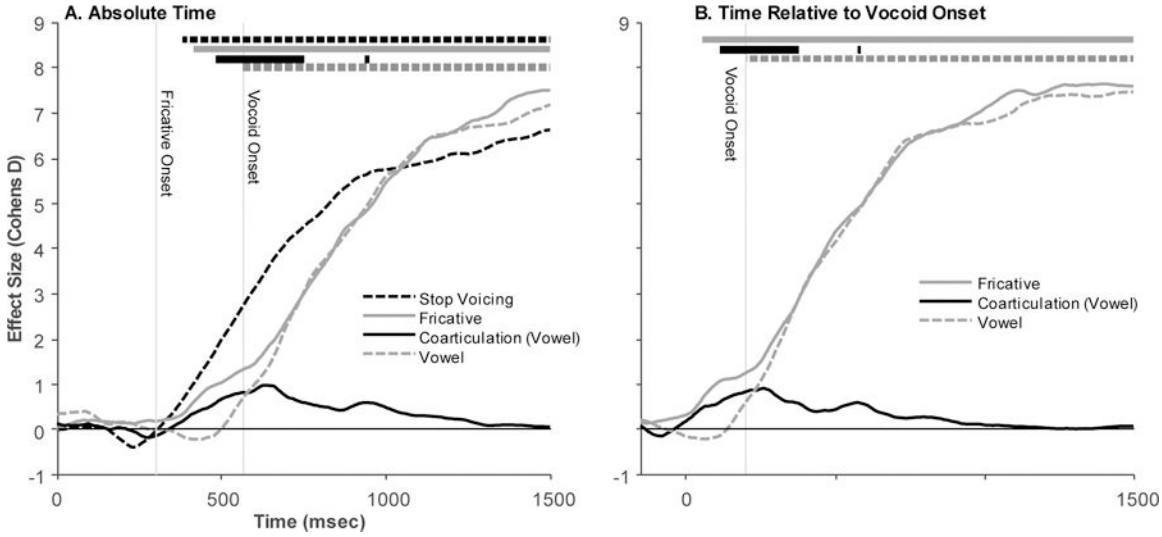

Figure 3.

Effect size (Cohen's D) of each of the four effects as a function of time. Vertical bars represent the onset of the fricative and the average onset of the vocoid in the stimulus, adjusted for the 200 msec oculomotor delay. Bars at the top of each panel indicate whether the corresponding effect was significantly different from 0 at that time (family-wise error corrected. A) As a function of time since trial onset. B) As a function of time since vocoid onset.

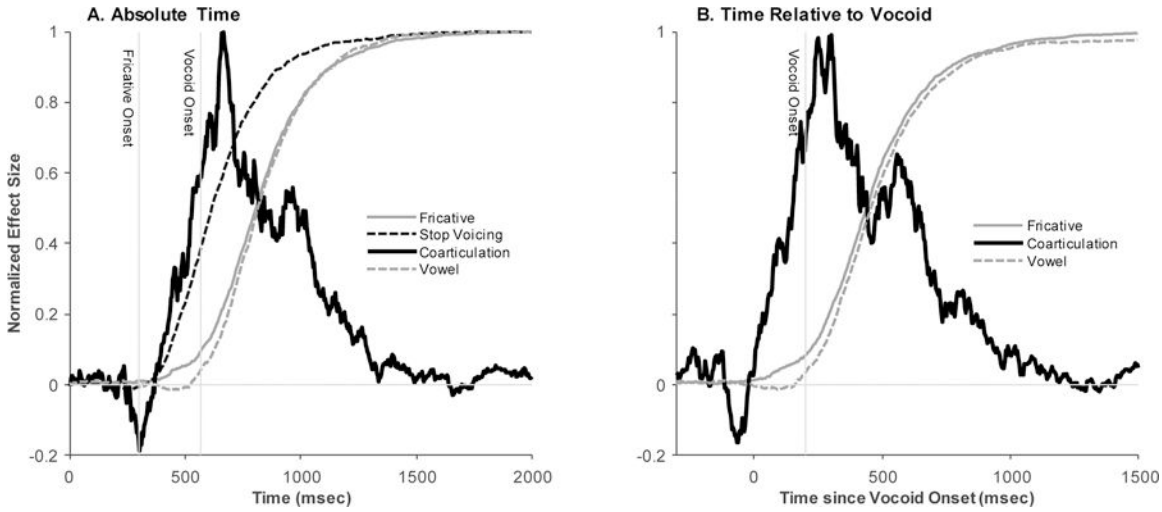

Figure 4.

Normalized effect size over time for fricative discrimination based on the frication spectrum (gray line), vowel discrimination based on the vocoid (black, dashed line), vowel discrimination based on anticipatory coarticulation ion the frication (black solid line), and stop voicing discrimination on the filler trials (black dotted line). A) As a function of time since trial onset. Here, 0 ms is the onset of the trial. Vertical lines are the fricative onset and mean vocoid onset (adjusted for the oculomotor delay). B) As a function of trial since the vocoid onset (Fricative trials only). Here, 0 ms is the onset of the vocoid (not adjusted for the oculomotor delay, thus the vertical line denoting vocoid onset is at 200 msec). Note that the coarticulation curves are noisier because they reflect a smaller absolute effect size, and are therefore more susceptible to small fluctuation when amplified by normalization.

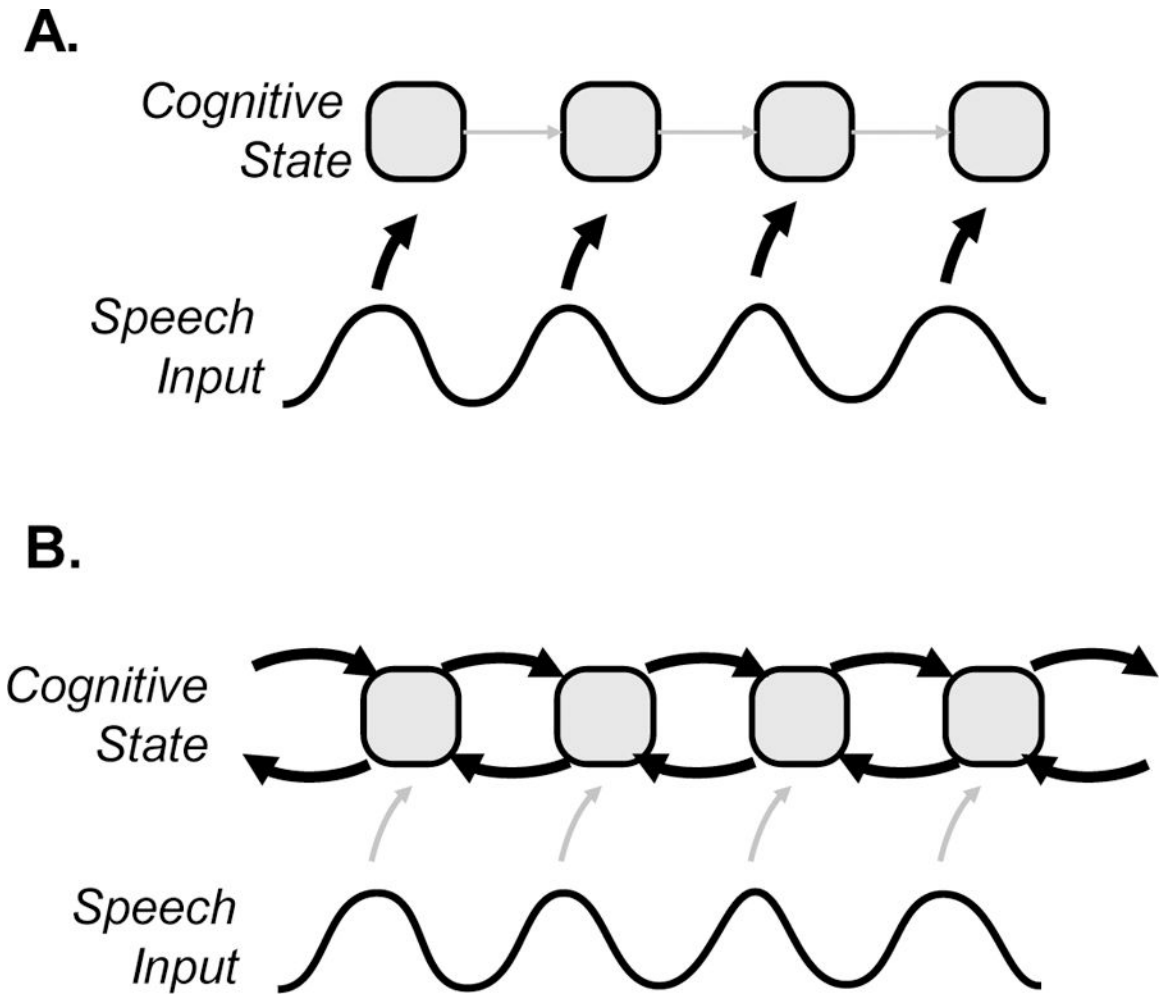

Figure 5.

Two hypothesized models relating time in the world to the unfolding cognitive state. A) In the standard models, the dynamics of the cognitive system are predominantly a function of the changing input; B) proposed here, the cognitive system has it's own dynamics which are modulated by the input.

Table 1:

Example grouping used in the first item-set.

Contrast	Context	Target	Competitor	Unrelated	Unrelated
Vowel Rounding	/s/	*Seed*	*soup*	*bite*	*bath*
	/ʃ/	*sheep*	*shoot*	*pit*	*pick*
Fricative Place	/i/	*Seed*	*sheep*	*bite*	*pit*
	/u/	*Soup*	*shoot*	*bath*	*pick*

Table 2:

Summary of significant regions for each effect. Auto-correlation (ρ) and corrected alpha (α*) are computed as part of familywise error correct. Left columns refer to time since trial onset. Right columns are relative to vocoid onset. Time regions are reported in ms and not adjusted for the 200 msec oculomotor delay.

	Absolute Time (Figure 3A)			Vocoid-Relative Time (Figure 3B)		
Effect	**ρ**	**α***	**Region**	**ρ**	**α***	**Region**
Stop	0.9955	0.00255	380 – 1500			
Fricative	0.9950	0.00247	416 – 1500	0.9954	0.00236	56 – 1500
Coarticulation	0.9977	0.00334	484 – 748	0.9980	0.00356	116 – 372
Vowel	0.9960	0.00270	560 – 1500	0.9961	0.00252	204 – 1500

Table 3:

Timing of stop voicing and fricative effects in fixation record relative to trial onset.

	Effect Onset (absolute time)		Stop vs. Fricative	
Threshold	***Stop Voicing (msec)***	***Fricative (msec)***	***t(30)***	***p***
20%	445	593	7.20	<.0001
30%	473	641	11.67	<.0001
40%	505	674	13.83	<.0001

Table 4:

Timing of fricative, transition, and rounding effects in eye-movement data. Time 0 is the onset of the vocoid.

	Effect Onset			Fricative vs. Vowel		Fricative vs. Coarticulation		Vowel vs. Coarticulation	
Threshold	***Fricative (msec)***	***Vowel (msec)***	***Coarticulation (msec)***	***t(30)***	***p***	***t(30)***	***p***	***t(30)***	***p***
20%	300	331	29	2.572	.015	4.402	<.001	5.209	<.001
30%	354	376	64	1.836	.076	7.313	<.001	8.491	<.001
40%	400	421	89	1.64	.111	9.876	<.001	12.440	<.001