Received: 1 March 2019 | Revised: 31 July 2019 | Accepted: 1 August 2019

DOI: 10.1002/wcs.1517

ADVANCED REVIEW

WIREs COGNITIVE SCIENCE WILEY

Acoustic context effects in speech perception

Christian Stilp

Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky

Correspondence
Christian Stilp, Department of Psychological and Brain Sciences, 317 Life Sciences Building, University of Louisville, Louisville, KY 40292.
Email: christian.stilp@louisville.edu

Abstract

The extreme acoustic variability of speech is well established, which makes the proficiency of human speech perception all the more impressive. Speech perception, like perception in any modality, is relative to context, and this provides a means to normalize the acoustic variability in the speech signal. Acoustic context effects in speech perception have been widely documented, but a clear understanding of how these effects relate to each other across stimuli, timescales, and acoustic domains is lacking. Here we review the influences that spectral context, temporal context, and spectrotemporal context have on speech perception. Studies are organized in terms of whether the context precedes the target (forward effects) or follows it (backward effects), and whether the context is adjacent to the target (proximal) or temporally removed from it (distal). Special cases where proximal and distal contexts have competing influences on perception are also considered. Across studies, a common theme emerges: acoustic differences between contexts and targets are perceptually magnified, producing contrast effects that facilitate perception of target sounds and words. This indicates enhanced sensitivity to changes in the acoustic environment, which maximizes the amount of potential information that can be transmitted to the perceiver.

This article is categorized under:

Linguistics > Language in Mind and Brain

Psychology > Perception and Psychophysics

KEYWORDS

context effects, speech categorization, speech perception

1 | INTRODUCTION

Humans hear speech more than any other sound. Our experience hearing speech starts in utero and persists across the lifespan. We are incredibly proficient at perceiving and understanding speech, so much so that it can be done in extremely challenging listening conditions (for reviews see Assmann & Summerfield, 2004; Mattys, Davis, Bradlow, & Scott, 2012). Yet, questions abound as to how we achieve such mastery, especially when considering the extreme acoustic variability in the speech signal (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952). Some have proposed that this variability can be overcome through certain invariant acoustic cues to speech sound identity (Blumstein & Stevens, 1979; Stevens & Blumstein, 1978), but it appears more likely that there are no invariant acoustic cues in speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967 and others). This returns us to our original question: how do we become so proficient in speech perception?

This problem is simplified if one returns to first principles about sensory and perceptual systems. Whatever the organism or the stimulus, all perception takes place in context. Stimulus variability becomes less problematic when one acknowledges that the perceiver is making judgments relative to context rather than based on absolute stimulus values. Given that sensory receptors have limited dynamic ranges that can only encode a fraction of the stimulus energy in the environment, this approach is as practical as it is efficient. Perceiving relative to context is a way to normalize stimulus-level variability, however extreme it may be. This point has a long but underappreciated history in speech perception. Acoustic properties within individual speech sounds (i.e., intrinsic cues) obviously contribute directly to their perception, but are highly variable. Acoustic properties of surrounding sounds form a vitally important context for perception (i.e., extrinsic cues; Joos, 1948; Ainsworth, 1975; Nearey, 1989), and they too contribute to perception of a given speech target. Listeners make use of both intrinsic and extrinsic cues to recognize speech sounds.

What exactly is meant by perceptual context? Context could span milliseconds, minutes, months, or even a lifetime of perceptual experience. For the purpose of this review, context is limited to only a few seconds before or after a given (target) sound. While this might sound restrictive initially, consider the wealth of acoustic information that speech contains on this timescale. For clues on how to delineate intrinsic versus extrinsic cues to speech sound identity, we turn to Repp's (1982) valuable distinction between trading relations and context effects. Trading relations are instances where speech sound perception maintains when changes in one contributing cue are offset by changes in another cue. Thus, trading relations occur among cues for the same speech sound contrast, which corresponds to intrinsic cues to speech sound identity. Context, on the other hand, is separate from direct cues for a given distinction and how it is produced; this corresponds to extrinsic cues to speech sound identity. To highlight this distinction, Repp cited fricative identification in fricative-vowel syllables as exhibiting a trading relation between frication noise and formant transitions and a context effect of the following vowel (cf., Mann & Repp, 1980). Given the continuous nature of the speech signal, universal agreement might never be achieved regarding which cues are purely intrinsic versus purely extrinsic or which situations constitute trading relations or context effects. Here we lean toward being slightly more inclusive in our review (such that a few of the studies reviewed below also appear in Repp's, 1982 review) as opposed to overly restrictive.

Dozens of studies have reported influences of surrounding acoustic context on perception of a given speech sound or word. Two theoretical accounts fiercely debate the origins of these context effects. One position has argued that speech production is fundamental to speech perception. Given the lack of one-to-one correspondence between speech acoustics and speech sound identity (e.g., Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Liberman et al., 1967), Motor Theory proposed that speech perception is accomplished through the recovery of intended articulatory gestures using a specialized speech-specific decoder (e.g., Liberman, 1996; Liberman et al., 1967; Liberman & Mattingly, 1985). Fowler (1986, 1996, 2006) proposed Direct Realism as an alternative to Motor Theory wherein the recovered objects of perception are actual articulatory gestures as opposed to intended ones. In this framework, perceivers need not be endowed with a specialized decoding mechanism to recover articulatory gestures, but instead perceive speech in an ecological framework consistent with that of Gibson (1979). Thus, perceptual resilience to acoustic variability wrought by coarticulation is provided by recovery of articulatory gestures—the neuromotor commands preceding them in Motor Theory, the actual gestures in Direct Realism.

The competing theoretical perspective has argued that articulatory gestures are neither necessary nor sufficient for producing acoustic context effects. Instead, this auditorist approach argues that context effects are merely byproducts of sufficiently developed auditory systems. Studies by Diehl, Kluender, Lotto, and Holt among others (reviewed below) have replicated acoustic context effects in speech perception using nonspeech contexts and/or target sounds (e.g., pure tones, music, noise). Collectively, these results advance a general auditory approach where differences in acoustic characteristics across context and target sounds were perceptually magnified, producing contrast effects. Contrast effects are not specific to human speech perception but are universal across all sensory modalities (von Békésy, 1967; Warren, 1985), thus requiring neither human perceivers nor speech stimuli to occur. The debate between gesturalist and auditorist approaches to speech perception has spanned decades and still continues (for reviews, see Fowler, Brown, & Mann, 2000; Diehl, Lotto, & Holt, 2004; Fowler, 2006; Lotto & Holt, 2006; Viswanathan, Fowler, & Magnuson, 2009; Viswanathan, Magnuson, & Fowler, 2010; Kingston et al., 2014; Rysling, Jesse, & Kingston, 2019). In the context of this review, general auditory mechanisms offer a more parsimonious explanation of speech and nonspeech perception by human and nonhuman listeners across shorter and longer timescales than do speech-production-specific accounts.

While there is little disagreement that context effects in speech perception occur, this literature suffers from several serious shortcomings. As reviewed below, context effects occur on various timescales, but individual studies typically focus on a single timescale. This makes it potentially perilous to propose that context effects on one (e.g., shorter) timescale support or predict context effects on a different (e.g., longer) timescale. Similar caution is warranted when context effects in one acoustic

domain (e.g., temporal) are taken to motivate or support inquiries in a different domain (e.g., spectral). Finally, researchers are quick to characterize the context effect in their studies but are often reticent to characterize the nature(s) of these context effects more broadly and/or provide a theoretical account.

Here we organize these vast literatures to clarify how acoustic context shapes speech perception. While this review might not be fully comprehensive, it is sufficiently detailed to reveal overarching trends in speech context effects on different timescales, in different directions, and in different acoustic domains (Figure 1). We characterize context effects in terms of whether the context is proximal (temporally adjacent to the target, generally shorter-duration) or distal (further displaced in time from the target, generally longer-duration), and whether the effects are forward (context precedes the target sound) or backward (target sound precedes the context). In a small number of cases, proximal and distal contexts are put in direct conflict with each other, making different predictions for perception of the target. The timescales and directions of context effects are reviewed in the spectral domain and the temporal domain, with two additional cases in the spectrotemporal domain also discussed.

2 | SPECTRAL CONTEXT EFFECTS

2.1 | Forward effects of proximal context

Spectral characteristics of surrounding sounds bias categorization of the target speech sound. Demonstrations of this phenomenon date back at least to Lindblom and Studdert-Kennedy (1967), who explored categorization of vowels /ɪ/ and /ʊ/ (as in the words "hid" and "hood," respectively) surrounded by either /w/ or /j/. The lower-frequency onset and offset of second-formant (F_2) transitions in /w/ contributed to more /wɪw/ (higher F_2) responses, and the higher-frequency onset and offset of F_2 in /j/ contributed to more /jʊj/ (lower F_2) responses. This finding was later extended to perception of medial vowels in /bVb/ and /dVd/ frames, where lower-F_2 /b/ produced more higher-F_2 /ɛ/ (as in "head") responses and higher-F_2 /d/ produced more lower-F_2 /ʌ/ (as in "hug") responses (Holt, Lotto, & Kluender, 2000; Nearey, 1989). In all of these cases, lower-frequency contexts biased categorizations toward higher-frequency vowels, and higher-frequency contexts biased responses toward lower-frequency vowels. Preceding and following contexts affected vowel categorization in these studies, producing results that are entirely consistent with the effects of preceding context reviewed next.

Mann (1980) reported that the preceding liquid consonant affected categorization of the subsequent stop consonant in /a/-consonant-consonant-/a/ (/aCCa/) stimuli: /d/ (higher F_3 onset) was perceived more often when following /r/ (lower F_3 offset), and /g/ (lower F_3 onset) was perceived more often when following /l/ (higher F_3 offset; see also Mann, 1986; Fowler, Best, & McRoberts, 1990). Similarly, the initial fricative in fricative-stop-vowel syllables exerted a contrastive influence on perception of the medial stop: higher-frequency /s/ promoted more /k/ identifications (lower-frequency formant transition onsets), and lower-frequency /ʃ/ (the consonant in "she") promoted more /t/ identifications (higher-frequency formant transition onsets; Mann & Repp, 1981).

Lotto, Kluender, and Holt (1997) replicated the /d/-/g/ categorization shifts reported by Mann et al. in nonhuman animal subjects (Japanese quail), strongly challenging the invocation of speech-production-specific underlying mechanisms. The necessity of articulatory gestures was further challenged in a series of studies by these authors where speech categorization shifts were observed when speech contexts were replaced with nonspeech (pure tone) stimuli (e.g., Lotto & Kluender, 1998; Holt & Kluender, 2000; Holt et al., 2000; see also Holt, 2005, 2006). In these cases, spectral differences across context and target sounds were perceptually magnified, producing what are known as spectral contrast effects (SCEs). SCEs were not exclusive to speech or even humans, but are instead produced through general operating characteristics of auditory systems.

However, this alternative explanation far from settled the debate. In some cases, these accounts were difficult to square because they made the same prediction for different reasons (e.g., more /d/ responses following /r/ contexts due to its low F_3 offset frequency [auditorist] or anterior place of articulation [gesturalist]). To distinguish this overlap, Viswanathan et al.

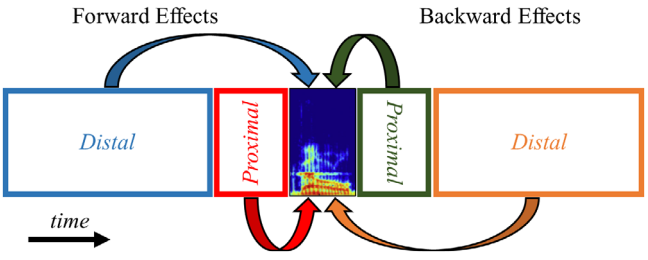

FIGURE 1 Acoustic context effects in speech perception. Contexts can be temporally adjacent to the target speech sound (proximal) or temporally nonadjacent to the target (distal). Contexts that precede the target in time are forward effects; contexts that follow the target are backward effects. These combinations of context timescales and directions apply equally to spectral context effects and temporal context effects in speech perception

(2010) tested compensation for coarticulation following liquid consonants in Tamil, a language where F_3 and place of articulation are distinct, thus making differing predictions for gesturalist and auditorist accounts of speech perception. Categorization of /da/-/ga/ targets followed articulation and not acoustics, offering support for gesturalism (but see Kingston et al., 2014 for an acoustic reinterpretation of this result).

A separate but related line of research has also documented enhanced processing of spectral differences over time. Consider a multitone complex where one (target) frequency has been removed. When this context precedes a nearly identical tone complex that does contain the target frequency, it becomes highly perceptually salient or "pops out," dramatically increasing its detectability. This is known as an auditory enhancement effect (EE), which has a long history in psychoacoustics research (Schouten, 1940; Viemeister, 1980; Viemeister & Bacon, 1982; and many others). With regard to speech perception, Summerfield and colleagues (Summerfield, Haggard, Foster, and Gray, 1984; Summerfield, Sidwell, and Nelson, 1987) leveraged EEs to make nonspeech stimuli sound like vowel sounds. In these studies, the target stimulus was a harmonic spectrum without any spectral peaks. The preceding contexts were also harmonic spectra, but with three narrow spectral notches at frequencies corresponding to vowel formant frequencies. When this context preceded the target stimulus, frequencies corresponding to vowel formants were perceptually enhanced, making the target stimulus sound like a vowel. Summerfield and Assmann (1987, 1989) demonstrated that EEs improved perceptual accuracy when contexts (an isolated vowel) and targets (a pair of concurrent vowels, one of which matched the context vowel) were both speech. The changes in spectral amplitude upon introduction of the second (noncontext) vowel in the target pair "popped out," making it far better recognized than when no precursor was presented. Later, Coady, Kluender, and Rhode (2003) presented harmonic spectra before /ba/-/da/ target stimuli. When contexts had spectral notches at frequencies appropriate for a low-F_2 vowel (like /o/), this enhanced the onset of the low-F_2 transition in the subsequent /b/, increasing listeners' /ba/ responses through EEs. Likewise, when contexts had notches at frequencies appropriate for a high-F_2 vowel (/e/, as in "way"), this enhanced the onset of the high-F_2 transition in /d/, increasing /da/ responses through EEs. Coady et al. also presented complementary contexts that only had energy at these notched frequencies, producing SCEs that biased responses in the other direction (low-F_2 contexts increased high-F_2 categorizations and vice versa).

2.2 | Backward effects of proximal context

Considerable attention has been paid to forward contexts effects (i.e., context precedes target), as this reflects the manner in which speech and all other acoustic signals unfold in time. A smaller but growing volume of work examines backward context effects, where contexts follow targets. Perhaps most well-known among backward context effects is Liberman, Delattre, and Cooper's (1952) investigation of unvoiced stop consonant perception. Categorization of the initial consonant based on its noise burst exhibited complex dependencies on the following vowel. Notably, perception of the exact same frequency burst changed as a function of following context (perceived as /p/ when preceding /i/ or /u/; perceived as /k/ when preceding /a/; also see Kiefte & Kluender, 2005). Other backward context effects in speech perception are more readily explained by spectral characteristics of the context, but are not uniform in the direction of their influence. In some cases, following contexts exerted a contrastive influence on preceding targets. In fricative-vowel syllables, the initial fricative was perceived as the lower-frequency /ʃ/ more often when followed by a higher-frequency vowel (/i/ as in "heat" or /a/ as in "hot"), but was perceived as the higher-frequency /s/ more often when followed by a lower-frequency vowel (/u/ as in "hoot"; Mann & Repp, 1980; Winn, Rhone, Chatterjee, & Idsardi, 2013). Watkins and Makin (1996) altered subsequent sounds' spectra in order to bias perception of the syllable-initial target vowel ("itch"-"etch," "apt"-"opt") or consonant ("slow"-"flow"). Contexts were processed by filters that accentuated spectral characteristics of one target continuum endpoint or the other (spectral envelope difference filters; Watkins, 1991). In all cases, responses were biased away from frequencies accentuated in the following contexts (e.g., accentuating frequencies for /f/ in the context produced more /s/ responses to the target).

Other investigations reported assimilative influences of context on the preceding target. Fujimura, Macchi, and Streeter (1978) examined perception of stop consonants (/b/, /d/, /g/) flanked by vowels (/a/, /e/, or /o/) in /VCV/ frames. Consonant identification assimilated more to the formant transitions into the final vowel (backward effect) than formant transitions from the initial vowel to the consonant (forward effect). Recently, Rysling et al. (2019) demonstrated the importance of temporal order for the nature of the context effect. In /hVC/ stimuli, unambiguous vowel contexts contrastively influenced categorization of the subsequent target consonant (higher-F_2 /i/ produced more lower-F_2-onset /p/ responses, lower-F_2 /u/ produced more higher-F_2-onset /t/ responses). When roles were reversed, unambiguous consonant contexts had assimilative influences on preceding vowel targets (more /u/ responses produced by following /p/ contexts, more /i/ responses produced by /t/ contexts). These patterns were replicated in /CV/ presentations, where consonant contexts (/b/, /d/) contrastively biased subsequent

vowel categorization (/i/, /u/), but consonant categorization assimilated to subsequent vowel contexts. Rysling et al. noted that spectral continuity between initial target and following context is a requirement for assimilation; otherwise, a spectrally discontinuous transition can result in a contrast effect (as in Mann & Repp, 1980; Winn et al., 2013).

In some cases, concurrent contrastive and assimilative effects have been reported in the same stimuli. Repp (1983) measured categorization of /b/ and /d/ in /aCCa/ stimuli with varying closure durations between syllables. The first syllable generally had a contrastive effect on identification of the consonant in the second syllable, but the second syllable had a variable influence on the first consonant depending on the closure duration. When closure durations were shorter, this effect was assimilative; when closure durations were longer, this effect was contrastive. This was replicated and extended by Wade and Holt (2005a), who presented /CVC/ stimuli with part of the medial vowel /a/ replaced by a low- or high-frequency pure tone. When the tone was presented immediately after formant transitions of the initial stop consonant, its influence was assimilative: 2800-Hz tones produced more /d/ (high-F_3) responses and 1800-Hz tones produced more /g/ (low-F_3) responses. When the tone was instead introduced 40 ms later in the vowel, its influence was instead contrastive. Elucidating the principles driving contrastive versus assimilative influences of following context remains an open and exciting area of research (for discussion see Rysling et al., 2019).

2.3 | Forward effects of distal context

In a seminal paper, Ladefoged and Broadbent (1957) demonstrated that the spectrum of a preceding sentence context influenced categorization of subsequent vowel targets. When F_1 frequencies in the context sentence ("Please say what this word is") were shifted higher, listeners were more likely to perceive the target word as "bit" (with low-F_1 /ɪ/). When F_1 frequencies in the context sentence were shifted lower, listeners were more likely to perceive the target word as "bet" (high-F_1 /ɛ/). The ensuing decades revealed the impressive generality of these distal SCEs, as speech categorization was biased when the preceding context: had a formant frequency shifted higher or lower (Ladefoged & Broadbent, 1957); had relatively narrowband (100 Hz; Stilp, Anderson, & Winn, 2015) to fairly broadband frequency regions amplified (1,000 Hz; Holt, 2006); was filtered by the difference between spectral envelopes of the target items (Watkins, 1991); was spoken with varying tongue and lip positions to alter formant frequencies (Ladefoged, 1989); or, was spoken by a talker with a much shorter vocal tract than s/he who spoke the target item (Dechovitz, 1977). Distal SCEs have biased speech targets differentiated by fundamental frequency (f0; Johnson, 1990), F_1 (Ladefoged & Broadbent, 1957), F_2 (Mitterer, 2006), F_3 (Holt, 2005), and overall spectral shape (Watkins, 1991; Watkins & Makin, 1996; see Stilp et al., 2015 for review). These findings are highly generalizable, as nonspeech contexts can bias phoneme categorization (signal-correlated noise: Watkins, 1991; pure tone sequences: Holt, 2005, 2006), and both speech and nonspeech contexts bias categorization of nonspeech stimuli (musical instruments; Stilp, Alexander, Kiefte, & Kluender, 2010). Recent work has shown that the magnitudes of distal SCEs are closely linked to the size of the spectral difference between context and target (in vowel categorization: Stilp et al., 2015; Stilp & Alexander, 2016; in consonant categorization: Stilp & Assgari, 2017; in musical instrument categorization: Frazier, Assgari, & Stilp, 2019). Importantly, these SCEs occur following more naturalistic context stimuli (unfiltered stimuli that already possess the desired spectral properties to produce SCEs), significantly enhancing their ecological validity (Lanning & Stilp, 2019; Stilp & Assgari, 2019).

Distal contexts have also produced EEs in speech perception. Like Coady et al. (2003), Holt (2006) presented spectrally complementary contexts before speech targets. When contexts were sequences of pure tones, this contrastively affected categorization of /da/-/ga/ targets via SCEs (higher-frequency tones produced more lower-F_3-onset /ga/ responses and vice versa). When contexts were wideband noise with spectrotemporal notches at the same frequencies as these pure tone sequences, consonant categorization shifted in the opposite direction via EEs (higher-frequency notches produced more higher-F_3-onset /da/ responses and vice versa). Stilp (2019) extended this framework to spectrally complementary sentence contexts. Sentences with notches in their spectra biased speech sound categorization through EEs, and sentences comprised of passbands at those same notch frequencies biased categorization through SCEs. These patterns were observed across both consonant categorization (/da/-/ga/) and vowel categorization (/ɪ/-/ɛ/). Additionally, EE magnitudes and SCE magnitudes in consonant categorization were significantly correlated with each other, suggesting general sensitivity to preceding context in these frequency regions.

Another example of distal spectral context effects is spectral calibration (also described as auditory perceptual calibration), which draws strong parallels to visual color constancy (Kiefte & Kluender, 2008). In spectral calibration, a spectral property that is perceptually salient for distinguishing speech sounds (e.g., F_2 frequency or overall spectral shape, both of which distinguish /i/ from /u/) is made reliable or predictable across the preceding context sentence and subsequent target sound, as though

imposed by the listening environment. In this situation, listeners decreased their reliance on the consistent spectral cue (e.g., context and target sounds sharing a spectral peak at the vowel's F_2 frequency) and increased their reliance on changing spectral cues that were more informative for speech sound identification (e.g., spectral tilt, which was varying throughout the trial). This occurs even when spectral peaks shared across context and target sounds were very modest (Stilp & Anderson, 2014). Additionally, while spectral calibration is highly sensitive to context characteristics such as its duration and sampling of frequency regions over time (Alexander & Kluender, 2010), it is relatively agnostic as to the source of the shared spectral properties (e.g., produced naturally or manipulated synthetically; Stilp, Anderson, Assgari, Ellis, & Zahorik, 2016). Alexander and Kluender (2010) argued that spectral calibration is closely related to SCEs, as the former involves calibrating to or deemphasizing unchanging stimulus properties (those that are shared across sounds) whereas the latter involves emphasizing changing stimulus properties (those that are different across sounds).

2.4 | Backward effects of distal context

Before his systematic investigation of distal forward context effects (Watkins, 1991), Watkins (1988, 1989) conducted parallel studies where the context followed the target vowel. A flat-spectrum sound was perceived as a vowel when the following context phrase was filtered by the inverse of the vowel's spectrum (Watkins, 1988; cf. Summerfield et al., 1987). Additionally, perception of /ɪ/ or /ɛ/ was biased when subsequent sounds in the trial ("/t/ is the next word") were filtered (Watkins, 1989). Watkins noted that these effects, while statistically significant, biased speech categorization to much smaller degrees than forwards effects did. To date, the clearest test of forward and backward distal spectral context effects was conducted by Sjerps, Zhang, and Peng (2018), who studied tone normalization and vowel normalization in Cantonese listeners. Tone perception was measured relative to a disyllabic context with raised or lowered f0 values, and vowel perception (/o/-/u/ target continuum) was measured relative to the same disyllabic context but with raised or lowered F_1 values. Sjerps et al. manipulated the trial structure (whether contexts preceded or followed targets) and the block structure (whether higher-frequency and lower-frequency contexts were tested in separate blocks or mixed within a single block). When contexts preceded targets, SCEs were observed in tone and vowel categorization for both blocked and mixed presentations, consistent with results reviewed earlier. However, when targets preceded contexts, results diverged. Tone categorization exhibited large SCEs in both blocked and mixed presentations, but vowel categorization exhibited SCEs only in blocked presentations. A follow-up analysis revealed that vowel categorization was being biased by the context on the *previous* trial, forming a "new" forward trial comprised of context from the preceding trial, a long interstimulus interval (the timing between two trials), and the target at the beginning of the subsequent trial. This result echoes a study by Broadbent and Ladefoged (1960), where SCEs were robust across 5-s intervals between contexts and vowel targets (and half of listeners exhibited SCEs across 10-s intervals), but backward context effects (vowel targets followed by the sentence context) were negligible.

2.5 | Proximal versus distal effects

The relative contributions of proximal and distal spectral context to speech perception have been considered by putting them in direct conflict. Holt (2006) tested pure tone contexts where their global (i.e., distal) properties (mean frequency across the full 2,100-ms duration = 2,300 Hz) could diverge from their local (i.e., proximal) properties (mean frequency of a 700-ms segment = 1,800, 2,300, or 2,800 Hz). Tone contexts consisted of three successive 700-ms epochs, each with a different mean, tested in all six possible orders. Statistical analyses failed to find consistent effects of these contexts on /da/-/ga/ categorization, leading Holt (2006) to conclude that local statistical structure (the epoch immediately preceding the target consonant) was far less perceptually salient than the global statistical structure. However, the use of analysis of variances across all six context conditions might have resulted in Type II error. Planned comparisons on proximal effects (e.g., sequences ending with mean frequency = 1,800 Hz [predicted to produce more "da" responses] vs. sequences ending with mean frequency = 2,800 Hz [predicted to produce more "ga" responses]) might have produced clearer results.

Later, Stilp (2018) tested this question by presenting sentence contexts before /da/-/ga/ targets. He measured the inherent balance of spectral energy across two frequency regions in sentences (low F_3 region: 1,700–2,700 Hz, high F_3 region: 2,700–3,700 Hz) using mean spectral differences (MSDs; Stilp & Assgari, 2019). MSDs were measured in two different temporal windows of the context sentences: the last 500 ms of the sentence (the Late window) and everything preceding the last 500 ms (the Early window). Context sentences were presented based on having competing MSDs at different points in time (e.g., Early window having more energy in low-F_3 frequencies, Late window having more energy in high-F_3 frequencies, and vice versa). Other sentences had approximately equal energy at low-F_3 and high-F_3 frequencies in the Early window and a

strong bias toward one frequency region in the Late window. In both conditions, (proximal) energy in the Late window predicted an SCE would occur, but the (distal) long-term balance of spectral energy was not strongly biased in either direction so no SCE was predicted to occur. Consonant categorization was biased via SCEs, revealing that the Late window of context sentences influenced responses despite ambiguous or even competing spectral information in the Early window.

3 | TEMPORAL CONTEXT EFFECTS

We now transition to cases where temporal characteristics of surrounding sounds inform speech perception. It bears mention that the theoretical debate surrounding spectral context effects discussed earlier (i.e., gesturalism vs. auditorism) applies equally to the source of temporal context effects, arguing whether listeners are recovering information about articulation rate or merely comparing the durations of acoustic events.

While we review these studies using the same organization as above, some lacked clear divisions as to which timescale (proximal vs. distal) and direction (forward vs. backward) were responsible for producing context effects. In Pickett and Decker (1960), the context "He was the (target) of the year" influenced perception of stop closures and the potential boundary between target words "topic" and "top pick." Slower speaking rates resulted in more "topic" percepts (shorter perceived stop closure), and faster rates resulted in more "top pick" percepts (longer perceived closure). Gottfried, Miller, and Payton (1990) showed that the speaking rate of the context "So (target) seems good" disambiguated some pairs of spectrally ambiguous vowels (slow speech increasing the perception of shorter vowel durations: more "bit" than "beet" responses and more "bet" than "bat" responses) but not others (no effects of rate on "bit"/"bet" responses). Whichever the underlying timescale and direction of these context effects, temporal characteristics of the surrounding context contrastively affected perception of target sounds and words.

3.1 | Forward effects of proximal context

Vowel duration affects the perception of voicing in the subsequent consonant. Denes (1955) reported that shorter vowels made the consonant sound longer, increasing the number of voiceless /s/ responses in /jus/ (as in "the use"); longer vowels made the consonant sound shorter, increasing the number of voiced /z/ responses in /juz/ (as in "to use"). Raphael (1972) generalized this finding to perception of word-final voicing in stops, fricatives, and consonant clusters in /CVC/ and /CVCC/ frames. This was extended to medial voicing by Port and Dalby (1982), where perception of the minimal pairs "dibber"/"dipper" and "digger"/"dicker" was informed by the duration of the first syllable (shorter duration produced more voiceless responses and vice versa). In these cases, intrinsic cues to voicing in the target sound were still influential (its own duration in Denes, 1955; medial closure duration in Port & Dalby, 1982), but they were supplemented by extrinsic temporal cues from the previous sound.

3.2 | Backward effects of proximal context

Backward temporal effects of following sounds on perception of preceding ones have been studied extensively. Perhaps best known among these effects is that duration of the following vowel influences perception of syllable-initial consonants varying in manner of articulation (Miller & Liberman, 1979). Longer vowel durations decreased the perceived duration of formant transitions in the initial consonant, resulting in more /b/ percepts; shorter vowel durations increased the perceived duration of these formant transitions, resulting in more /w/ percepts. This effect is apparent in speech perception by infants (Miller & Eimas, 1980) and has been replicated using nonspeech stimuli (Diehl & Walsh, 1989; Pisoni, Carrell, & Gans, 1983), suggesting a very general auditory basis (but see Shinn, Blumstein, & Jongman, 1985; Miller & Wayland, 1993; Utman, 1998 for discussions of the prominence and replicability of this effect). Similar results have been reported for syllable-initial consonants varying in voice onset time (VOT), where longer subsequent vowel durations produced more short-VOT (voiced) percepts and shorter vowel durations produced more long-VOT (voiceless) percepts (e.g., Green & Miller, 1985; McMurray, Clayards, Tanenhaus, & Aslin, 2008; Miller & Dexter, 1988; Summerfield, 1981; Toscano & McMurray, 2012). This effect is not limited to shifting category boundaries between phonemes, but also shapes the internal structures of the categories (Miller & Volaitis, 1989; Volaitis & Miller, 1992). Finally, distinctions between syllable-initial fricatives and affricates are also informed by the duration of subsequent speech sounds. Longer vowel durations made the initial consonant sound shorter, producing more affricate responses (/tʃæs/, or "chass"); shorter vowels made the consonant sound longer, producing more fricative responses (/ʃæs/, or "shass"; Newman & Sawusch, 1996).

3.3 | Forward effects of distal context

Studies of temporal context effects invite particularly close consideration of timescale. Debate regarding what constitutes a proximal context has been minimal, but vast interpretation is available in what constitutes a distal context. Heffner, Newman, and Idsardi (2017) conducted a series of experiments where they explicitly varied the definition of distal context, from more than 400 ms before the target item (as in Newman & Sawusch, 1996), more than one syllable before the target item (as in Dilley & Pitt, 2010), or even the window between these two reference points. Influences of distal speaking rate were not completely consistent across segment perception and word segmentation experiments, highlighting the importance of precision when defining temporal context. It is difficult to envision any single cutoff point adequately delineating proximal from distal for all stimuli. Therefore, here the term is intentionally used broadly as to encompass the most studies and observe patterns of results therein.

Temporal distinctions between consonants reviewed in Section 2.2 are also shaped by distal forwards (preceding) context. A host of studies examined how context sentences spoken at slower rates resulted in perception of shorter VOTs (more voiced responses) and faster sentences resulted in perception of longer VOTs (more voiceless responses). These rate effects were reported for consonants in both initial (/b/-/p/: Wayland, Miller, & Volaitis, 1994; /g/-/k/: Diehl, Souther, & Convis, 1980; Summerfield, 1981; Kidd, 1989) and medial positions (/b/-/p/: Port, 1979; Port & Dalby, 1982; Gordon, 1988). In addition to shifting VOT category boundaries, context rate influenced internal category structure and locations of the "best" exemplars (Wayland et al., 1994). Similar rate effects have been reported for perception of /b/ and /w/, where faster context sentences shifted the category boundary to shorter formant transition durations (and more /w/ responses; Minifie, Kuhl, & Stecher, 1977; see Wade & Holt, 2005b for a replication using pure tone contexts). Repp, Liberman, Eccardt, and Pesetsky (1978) measured perception of "shop" and "chop" targets preceded by the context sentence "Why don't we say (target) again." In addition to intrinsic influences of silence duration and frication duration on consonant recognition, slow context sentences increased the number of (shorter-duration "ch") affricate response and fast context sentences increased the number of (longer-duration "sh") fricative responses.

Distal preceding contexts also inform vowel perception. Ainsworth (1972, 1974) demonstrated context effects produced by the rhythm of a sequence of isochronous /ə/ vowels (as in the first vowel in "about"). When this vowel sequence had a slower rhythm, participants perceived more short vowels (e.g., /ɪ/, /ɛ/, /ʌ/, and /ʊ/); when the vowel sequence had a faster rhythm, more long vowels were perceived (e.g., /u/, /i/, and /ɝ/ as in "heard"). However, effects were relatively confined to vowels whose formant frequencies were perceptually ambiguous. Recent work capitalized on a vowel contrast for which duration information is used contrastively: Dutch /ɑ/-/a:/. This vowel pair is distinguished both spectrally and temporally (higher F_2 frequencies and longer durations in /a:/), which has allowed for experiments that examine spectral and temporal context effects concurrently (Reinisch & Sjerps, 2013). As for distal effects of temporal context, the expected rate normalization effect occurs where faster preceding sentences produced more longer-duration /a:/ percepts and slower sentences produced more shorter-duration /ɑ/ percepts (Bosker, 2017a, 2017b; Maslowski, Meyer, & Bosker, 2018, 2019; Reinisch, 2016; Reinisch & Sjerps, 2013). This work has introduced a new and more ecologically valid experimental paradigm termed "habitual rate tracking," where context effects are observed following extended exposure to a talker in conversation (≈2 min) rather than one sentence at a time on a trial-by-trial basis (Maslowski et al., 2019; Reinisch, 2016). Potential asymmetries in characteristics of these rate normalization effects (whether a foreign language sounds faster than one's native language [Bosker & Reinisch, 2017]; whether one's own speech produces similar context effects as other talkers' speech [Bosker, 2017b; Maslowski et al., 2018]) highlight interesting avenues for further inquiry.

Finally, distal speech rate also affects word segmentation. Dilley and Pitt (2010) observed that speaking rate modulated listeners' perception of function words (e.g., or, are, a) in sentences that were grammatically acceptable with or without them. When the target phrase containing the function word was sped up or the rest of the sentence was slowed down, fewer function words were detected. When the target phrase without a function word was slowed down or the rest of the sentence was sped up, more function words were detected. This is known as the Lexical Rate Effect, where relative speaking rate influenced perception of function words in the target phrases. This effect builds as listeners accumulate experience with speaking rate information (Baese-Berk et al., 2014), interacts with other acoustic cues to a function word's presence (Heffner, Dilley, McAuley, & Pitt, 2013), and extends to perception of prosodically weak syllables more generally (Baese-Berk, Dilley, Henry, Vinke, & Banzina, 2019; see also Reinisch, Jesse, & McQueen, 2011a, 2011b for effects of distal rate on perception of prosody and word boundaries in Dutch). Interestingly, the Lexical Rate Effect appears to be specific to intelligible speech contexts, as spectrally degraded sentences and nonspeech tones failed to elicit changes in function word recognition (Pitt, Szostak, & Dilley, 2016).

3.4 | Backward effects of distal context

The duration of the following phoneme informs categorization of its antecedent, but this influence is very temporally limited. In Miller and Liberman's (1979) report on vowel duration influencing categorization of the preceding consonant, they extended their /CV/ stimuli to /CVCV/ (with the second syllable being /da/) to compare how the duration of each syllable influenced responses. When duration of the first syllable was short (80 ms), the second syllable exerted its own context effect where shorter durations (72 ms) produced more /w/ responses than longer durations (216 ms). However, when the duration of the first syllable was longer (224 ms), the second syllable had a negligible effect on responses. Then, they added 36-ms formant transitions appropriate for /d/ to the end of /ba/-/wa/ syllables to vary their perceived speaking rate. Relative to an 80-ms /ba/-/wa/ series, /b/ responses increased when syllable duration was lengthened to 116 ms (slower perceived rate) but decreased sharply when /d/ was appended to create 116-ms /Cad/ syllables (faster perceived rate).

Subsequent investigations further delimited the influence of later speech sounds on perception of earlier ones. The VOT boundary between /bi/ and /pi/ ("bee" and "pea") was shortened by similar amounts when various stimuli were appended to them (syllable-final /z/ with duration 37.5 or 87.5 ms; the word "again" with duration 375 or 560 ms; Summerfield, 1981). Newman and Sawusch (1996) and Sawusch and Newman (2000) suggested that phonemes only within approximately 300 ms of the earlier target phoneme influenced its perception. This pattern was first observed in syllable-initial /tʃ/-/ʃ/ categorization when the perceptually relevant proximal context was a vowel (/Cæs/), semivowel, (/Cwæs/), or stop (/Ckas/): duration of the proximal context modified fricative/affricate responses, but duration of the subsequent vowel did not. When it was of sufficiently short duration, then this proximal context could contain multiple speech sounds (durations of /l/ and /o/ biasing categorization of the initial consonant in /blos/-/plos/ and /dlos/-/tlos/ continua; Newman & Sawusch, 1996) or extend into the final sound in the test syllable (/bʊʃ/-/pʊʃ/ or "bush"-"push" continua; Sawusch & Newman, 2000).

3.5 | Proximal versus distal effects

Few studies of spectral context effects evaluated influences of proximal versus distal contexts on speech perception, but several studies of temporal context effects have. In most of these investigations, both distal and proximal contexts preceded the target item. Summerfield (1981) manipulated individual words in the context "why are you" so that each had a fast duration (110 ms) or slow duration (220 ms). Contexts preceded target items varying in the VOT of their initial stop consonant (/biz/-/piz/, or "bees"-"peas"). VOT boundaries shifted by 10 ms across the all-slow and all-fast contexts, but more telling were duration manipulations of individual words in the context. Varying the duration of only the first or only the second word shifted VOT boundaries by approximately 1–2 ms, but varying the duration of "you" shifted VOT boundaries by 6–7 ms, suggesting a much larger influence of proximal rate than distal rate on VOT perception. Similarly, Kidd (1989) manipulated word durations in the context sentence "A bird in the hand is worth two in the" before presenting /gi/-/ki/ ("gee"-"kee") targets. As expected, fast contexts produced shorter-duration VOT boundaries (more voiceless "kee" responses) than slow contexts. When word durations alternated in a regular rhythm, longer-duration stressed syllables produced similar results to all-slow sentences and shorter-duration stressed syllables produced similar results to all-fast sentences (but VOT boundaries were not quite as extreme as they were for all-slow/all-fast sentences). When these rhythmic patterns were violated, categorization was heavily influenced by proximal timing information. Despite several syllables' worth of evidence that speaking rate was slow, the fast duration of the last syllable produced short VOT boundaries (even shorter than boundaries produced by all-fast sentences), and vice versa. Later, Reinisch et al. (2011b) varied rate characteristics of context sentences before words that did or did not start with the target consonant (/s/-initial vs. non-/s/-initial; /t/-initial vs. non-/t/-initial). Both distal and proximal rate affected listeners' responses (faster rates produced more responses including the target consonant), but to different degrees: proximal rate had a larger effect on performance, which was only modestly attenuated by competing distal rate. When distal rate was slow/fast but the proximal rate was neutral, then distal rate biased responses. These relationships extend to perception of function words: when distal and proximal rate information conflict, proximal rate exerts a stronger influence on perception (as evidenced in larger regression coefficients for proximal cues than distal cues; Heffner et al., 2013). Finally, Reinisch (2016) tested the limits of the habitual rate tracking paradigm by using the extended conversational exposure as distal speaking rate information and context sentences on each trial as proximal rate information. Consistent with the different paradigms outlined above, proximal rate information had a much stronger influence on speech perception than earlier distal rate information.

How does speech perception resolve temporal context that precedes and follows the target item? Toscano and McMurray (2015) asked this question in perception of syllable-initial voicing in /b/-/p/. The distal forwards effect of speaking rate biased VOT perception (with faster rates producing more voiceless percepts, cf. Summerfield, 1981 and others), and the proximal

backward effect of subsequent vowel length also informed perception (with shorter vowels also producing more voiceless percepts, cf. Miller & Liberman, 1979 and others). When both contexts are present, one might predict that speaking rate would exert a larger effect on VOT perception than vowel length owing to it preceding the target item. Instead, speaking rate and vowel length exerted similar degrees of influence on voicing perception. Eyetracking data revealed that listeners did not immediately shift their gaze once they heard (at least some of) the context sentence rate; instead, looks to target items were delayed until they heard the VOT of the target item. The effect of speaking rate occurred when it could be compared to (and bias perception of) VOT; looks in response to vowel length information occurred later in the trial.

4 | SPECTROTEMPORAL CONTEXT EFFECTS

The aforementioned studies revealed effects of spectral or temporal context on perception of target sounds and words, but acoustic context in everyday listening is more spectrotemporal than it is purely spectral or purely temporal. Here we review two examples of how speech perception adjusts to spectrotemporal properties of the listening context. Unlike the diversity of effects reviewed above, these spectrotemporal context effects are forwards and distal.

Reinisch and Sjerps (2013) measured spectral and temporal context effects simultaneously in the perception of Dutch vowels /ɑ/-/a:/, which are distinguished both spectrally and temporally. This permitted manipulation of both spectral (lower or higher F_2) and temporal (slower or faster speaking rates) characteristics of preceding context sentences. Eyetracking data revealed that the context spectrum biased listeners' looks to the target item far earlier than the context rate did. Similar to Toscano and McMurray (2015), looks in response to context rate did not occur until the target was presented so that the relative rates could be compared. When considering intrinsic characteristics of the target vowel, its spectrum influenced listeners' looks slightly earlier than did the vowel's duration. Thus, spectral and temporal context effects both occurred, but with spectral effects slightly preceding temporal effects.

A given pattern of reverberation introduces spectrotemporal alterations to the source signal (often depicted using a room impulse response). It has long been known that reverberation can degrade speech intelligibility and/or quality, particularly when reverberation times are long (Knudsen, 1929). But, with sufficient exposure, listeners can compensate for reverberation and speech perception recovers. Therefore, while not traditionally viewed as such, reverberation is a spectrotemporal context to which perception can adjust. In a series of studies by Watkins and colleagues (Watkins, 2005a; 2005b; Watkins & Makin, 2007; Watkins, Raimond, & Makin, 2011) explored compensation for reverberation in word perception. They tested a series of target words varying from "sir" to "stir" in the degree of modulation in the amplitude envelopes (high modulation depth in "stir," low modulation depth in "sir"). In nonreverberant conditions, "stir" stimuli were perceived as such owing to the silent closure interval preceding the /t/. When reverberation was added to the word, spectral energy filled this silent gap and "sir" responses increased. When this reverberant target word was presented in a context sentence ("Next you'll get (target) to click on") with the same reverberation, perception compensated for the reverberation and "stir" responses increased accordingly (comparable to responses to this stimulus in nonreverberant conditions). Later, Zahorik and colleagues (Brandewie & Zahorik, 2010, 2013; Srinivasan & Zahorik, 2013; Zahorik & Brandewie, 2016) extended these effects to sentence intelligibility, as prior exposure to reverberation characteristics of a given room enhanced the intelligibility of speech heard in that room. From this perspective, compensating for reverberation is another instance of perceptual constancy in speech perception (Assmann & Summerfield, 2004; Watkins & Makin, 2007).

5 | DISCUSSION

Dozens of investigations spanning over a half-century have examined acoustic context effects in speech perception. With widely varying stimuli and approaches, it is remarkable that the results of these studies cohere at all. Across studies, three patterns emerge. First, forwards context effects generally exert larger influences on speech perception than backward effects. Second, effects of proximal context are generally larger than effects of distal context. Third and most importantly, in the vast majority of cases reviewed above, across forward and backward effects of proximal or distal context, whether the context was spectral, temporal, or spectrotemporal, a persistent theme emerges: contrast. Perception of the target sound or word was facilitated through the exaggeration of acoustic differences between it and surrounding sounds. This is no accident; it is a savvy processing strategy not just in speech perception but all of perception most broadly. Sensorineural systems respond primarily to change, so exaggerating differences via contrast effects makes new stimuli more perceptually salient (Kluender, Coady, & Kiefte, 2003; von Békésy, 1967; Warren, 1985). These principles are so formative that one can conceptualize a considerable

portion of speech perception as being change detection (Winn & Stilp, 2019). Kluender and colleagues (Kluender & Alexander, 2007; Kluender & Kiefte, 2006; Kluender, Stilp, & Kiefte, 2013; Kluender, Stilp, & Llanos, 2019) have promoted an information-theoretic approach to speech perception because emphasizing what is changing or unpredictable maximizes the potential information that can be transmitted to the perceiver. Many acoustic context effects are perfect examples of how acoustic differences between sounds are perceptually exaggerated to facilitate speech perception and ultimately guide adaptive behavior.

The influence of acoustic context on speech perception is broad, but it is certainly not the only acoustic influence on speech perception. Intrinsic acoustic cues to speech sound identity are not always unambiguous, but they often narrow the list of potential candidates to a manageable few. When intrinsic acoustic properties to speech sound identity are clear, the disambiguating influence of extrinsic acoustic context is lessened. In many of the investigations reviewed above, target stimuli were speech sound continua that transitioned from one clear endpoint to another. In such paradigms, context effects are most evident in categorization of mid-continuum members but not always the endpoints. Some might cite this as a limitation of the influence of acoustic context, but acoustic characteristics of fluent coarticulated speech often fall short of such extremes (Lindblom, 1963). As such, these hypoarticulated mid-continuum stimuli are generally more representative of the speech produced in everyday conversation, making context effects an important contributor to everyday speech perception.

Of course, speech perception is shaped by factors beyond intrinsic and extrinsic acoustic properties. Extensive review of such influences is beyond the scope of the present submission, but here we briefly note five such factors whose relationships and interactions with acoustic context effects have been considered. First, while this review extensively examines native-language speech perception, what role does nonnative language play in acoustic context effects? Sjerps and Smiljanic (2013) presented English, Dutch, and Spanish sentences before target items varying from /o/-/u/ to English, Dutch, and Spanish (monolingual and Spanish–English bilingual) listeners. Whatever the stimulus language or listener language background, low-F_1-amplified context sentences produced more /o/ response and high-F_1-amplified sentences produced more /u/ responses. These results imply that language background plays no role in spectral context effects, but this might only be true when all languages share the speech sound contrast under study (as was the case in Sjerps & Smiljanic, 2013). Kang, Johnson, and Finley (2016) asked native English and French listeners to categorize the initial consonant in fricative-vowel syllables. Fricative categorization was biased by the following vowel context (/a/ produced more /ʃ/ responses and /u/ produced more /s/ responses, replicating Mann & Repp, 1980). They tested a third vowel context, the rounded vowel /y/, which was familiar to French listeners but unfamiliar to English listeners. French listeners produced a unique pattern of responses to this third vowel context (responding differentially to /a/, /u/, and /y/ contexts) but English listeners did not. While this and related findings suggest a degree of language specificity, context effects have also been observed for speech sounds absent in a native language but present in a nonnative language. Mann (1986) reported that native Japanese listeners who could not distinguish English /l/ and /r/ nonetheless exhibited effects of preceding spectral context that mirrored English listeners (more /g/ responses following /l/ context, more /d/ responses following /r/ context; cf. Mann, 1980). Thus, there may exist complex interactions between the speech sound inventory in one's native language, the inventory in a particular nonnative language being tested, and acoustic context effects.

Second, lexicality shapes perception of speech sounds and words. Categorization of initial stop consonants is biased toward the response option that forms a valid word (more "dash" than "tash" responses and more "task" than "dask" responses; Ganong, 1980). Additionally, speech sound categorizations will increase in a particular direction if listeners previously heard words ending in that speech sound (Norris, McQueen, & Cutler, 2003). The intersection between these lexicality effects and acoustic context effects have been explored in both the spectral and temporal domains. Sjerps and Reinisch (2015) filtered preceding contexts in order to disambiguate a perceptually ambiguous sound in the target word as /s/ or /f/. When the target sound was perceived as the option consistent with a legal word, lexicality effects (from lexically guided perceptual learning; Norris et al., 2003) were eliminated. When the target sound was perceived as the other response option which did not make a valid word, lexical effects influenced categorization. In the temporal domain, Miller and Dexter (1988) replicated the influence of lexicality on VOT categorization where responses were biased toward valid words (cf. Ganong, 1980). Under speeded responding conditions, however, lexical effects were largely extinguished but temporal effects of speaking rate still influenced perception. Across both studies, acoustic context effects were resolved before lexicality effects could take place, consistent with the lower-level and higher-level nature of these respective influences on speech perception.

Third, listening to various different talkers is more challenging than listening to a single talker (Creelman, 1957; Mullennix, Pisoni, & Martin, 1989; Magnuson & Nusbaum, 2007; and many others). While SCEs in speech perception have been reported following nonspeech contexts (Holt, 2006; Watkins, 1991), these effects are sensitive to who spoke the context sentences. Assgari and Stilp (2015) reported that SCEs biasing vowel categorization were smaller when context sentences

were spoken by 200 different talkers (a new talker on each trial) as compared to just one talker. Consequences of talker variability seem to be related to acoustic similarity between talkers' voices: contexts spoken by different talkers with similar fundamental frequencies produced larger SCEs than different context talkers with highly variable fundamental frequencies (Assgari, Theodore, & Stilp, 2019).

Fourth, expectations can modulate speech perception, such as how many talkers are heard (Magnuson & Nusbaum, 2007) or which regional dialect a talker has (Hay & Drager, 2010). Johnson, Strand, and D'Imperio (1999) instructed their listeners to imagine that the target syllables varying from /hʊd/–/hʌd/ were spoken by a man or a woman. Listeners who were told the talker was a man exhibited a lower-frequency boundary between the target vowels than the listeners who were told the talker was a woman despite hearing the exact same stimuli. Acoustic characteristics of the gender-ambiguous stimuli were interpreted relative to listeners' expectations and experience with men's and women's voices.

Finally, visual cues are also important in speech perception. With regard to acoustic context effects, Glidden and Assmann (2004) compared the influences of fundamental frequency, spectral envelope, and visual talker gender on shifts in category boundaries between /ɪ/ and /ɛ/, which are principally distinguished by F_1 frequency. For each combination of spectral properties tested, seeing a female talker produced higher category boundaries than seeing a male talker (see also Johnson et al., 1999). Thus, visual information about the talker also forms a context for interpreting and categorizing speech sounds. Across these examples, contrast (the typical direction of these context effects) in no way accounts for all of speech perception, but many influences are integrated in order to perceive effectively in many different types of context (Lotto & Holt, 2006).

Given the diversity of acoustic context effects reviewed here and the variety of additional influences upon them, it should come as no surprise that a range of proposed mechanisms is likely at play. Many of these mechanisms have not yet been definitively confirmed, and some are still fiercely debated. An extensive examination of proposed mechanisms is worthy of detailed review in its own right, but here we will touch on some of these mechanisms at relatively lower and higher levels of processing, noting the context effect(s) that have been suggested to arise from them.

Physiological responses as simple as neural adaptation may contribute substantially to speech perception (e.g., Delgutte, 1996; Delgutte, Hammond, Kalluri, Litvak, & Cariani, 1996). In the presence of continued stimulation, neurons typically decrease their firing rates. This is not a weakness but a sophisticated response strategy that codes the fact that the stimulus is unchanging while also retaining sensitivity to when inputs change, that is, when new information is available (see Wark, Lundstrom, & Fairhall, 2007; Winn & Stilp, 2019 for discussions). Neurons adapted by earlier (context) sounds would be less responsive to those frequencies in subsequent (target) sounds; neurons that are unadapted/less adapted by earlier sounds would be relatively more responsive to their frequencies in later sounds, producing a shift in perceived frequency (SCEs: e.g., Delgutte, 1996; Delgutte et al., 1996; Holt et al., 2000; spectral calibration: Alexander & Kluender, 2010). Closely related is the adaptation of inhibition (also termed adaptation of suppression), where activated neurons also inhibit the responses of other neurons encoding adjacent frequencies. This inhibition adapts over time, resulting in neural responses to inhibited frequencies being more pronounced later (in response to the target) than they were initially (in response to the context), increasing their perceptual salience (EEs: e.g., Viemeister & Bacon, 1982; Summerfield et al., 1984; Nelson & Young, 2010). A third possible physiological mechanism influencing context effects is backward masking, where the following context impedes perception of the preceding target (often due to spectral overlap between the two; e.g., Pickett, 1959). This may be a candidate mechanism for backward contrast effects (such as those reported by Mann & Repp, 1980 and/or Watkins & Makin, 1996), but the large individual differences in backward masking requires targeted research to test this possibility.

Various higher-level mechanisms have also been proposed to contribute to context effects in speech perception. Perceptual grouping of similar stimuli and segregation of dissimilar stimuli is pervasive in auditory perception (Bregman, 1990). These operations have been suggested to contribute to a number of acoustic context effects (e.g., EEs: Darwin, 1984; Kidd & Wright, 1994; spectral calibration: Alexander & Kluender, 2010; backward spectral assimilation: Rysling et al., 2019). Also, recent work has revealed cortical entrainment to oscillations in the amplitude envelope of speech (e.g., Giraud & Poeppel, 2012; Peelle & Davis, 2012), and this entrainment has been proposed to contribute to speaking rate normalization (Bosker & Ghitza, 2018). Finally, the recovery of articulatory gestures proposed by Motor Theory (Liberman et al., 1967; Liberman & Mattingly, 1985) and Direct Realism (Fowler, 1986) has been proposed to explain various context effects including compensation for coarticulation (i.e., recovering gestures to undo effects of coarticulation; proximal spectral context effects) and articulation rate (accessing information about a talker's speaking rate; proximal and distal temporal context effects). This recovery would presumably occur cortically (e.g., involving premotor cortex for intended articulatory gestures and motor cortex for actual articulatory gestures, respectively), as precortical neural architecture is ill-equipped to do so. Given that the occurrences of acoustic context effects are well-instantiated in the literature, future research would be well served by focusing on establishing their underlying mechanisms.

6 | CONCLUSION

In speech perception, as with perception in general, context is essential. Context takes many different forms, exerting a multitude of potential influences on speech perception. Here, we examined acoustic contexts that spanned narrow (a pure tone or formant peak) to broad spectral bandwidths (spectral envelope) and narrow (tens of milliseconds) to broad temporal extents (several seconds). We reviewed influences of acoustic contexts that preceded or followed the target item, were adjacent to or temporally removed from the target item, or even competed with itself on different timescales. By and large, acoustic context serves to disambiguate target speech via contrast. Differences between context and target are perceptually magnified, producing contrast effects that facilitate speech perception. By doing so, sensitivity to changes in frequency and time is enhanced in order to maximize the amount of information that can be transmitted from environment to perceiver.

ACKNOWLEDGMENTS

We thank Joseph Toscano and an anonymous reviewer for extremely helpful feedback and suggestions on an earlier draft of this review.

CONFLICT OF INTEREST

The author has declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Early recognition of speech
Psychology of auditory perception
Speech perception and production

ORCID

Christian Stilp https://orcid.org/0000-0002-5119-201X

REFERENCES

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *The Journal of the Acoustical Society of America*, *51*(2), 648–651.

Ainsworth, W. A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, *17*(2), 103–109.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 103–113). London, England: Academic Press.

Alexander, J. M., & Kluender, K. R. (2010). Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context. *Journal of the Acoustical Society of America*, *128*(6), 3597–3613.

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *Journal of the Acoustical Society of America*, *138*(5), 3023–3032.

Assgari, A. A., Theodore, R. M., & Stilp, C. E. (2019). Variability in talkers' fundamental frequencies shapes context effects in speech perception. *Journal of the Acoustical Society of America*, *145*(3), 1443–1454.

Assmann, P. F., & Summerfield, Q. (2004). The perception of speech under adverse conditions. In *Speech processing in the auditory system* (Vol. 18). New York, NY: Springer.

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & Banzina, E. (2019). Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables. *Attention, Perception, & Psychophysics*, *81*(2), 571–589.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science*, *25*(8), 1546–1553.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production - Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Bosker, H. R. (2017a). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, *79*(1), 333–343.

Bosker, H. R. (2017b). How our own speech rate influences our perception of others. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1225–1238.

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: Behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, *33*(8), 955–967.

Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology*, *8*, 1–13. https://doi.org/10.3389/fpsyg.2017.01063

Brandewie, E., & Zahorik, P. (2010). Prior listening in rooms improves speech intelligibility. *Journal of the Acoustical Society of America*, *128*(1), 291–299.

Brandewie, E., & Zahorik, P. (2013). Time course of a perceptual enhancement effect for noise-masked speech in reverberant environments. *Journal of the Acoustical Society of America*, *134*(2), EL265–EL270.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgements and adaptation level. *Proceedings of the Royal Society B: Biological Sciences*, *151*, 384–399.

Coady, J. A., Kluender, K. R., & Rhode, W. S. (2003). Effects of contrast between onsets of speech and other complex spectra. *Journal of the Acoustical Society of America*, *114*(4), 2225–2235.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, *24*, 597–606.

Creelman, C. D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, *29*(5), 655.

Darwin, C. J. (1984). Auditory processing and speech perception. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 197–209). London, England: Erlbaum.

Dechovitz, D. (1977). Information conveyed by vowels: A confirmation. *Haskins Lab Status Report*, *SR-51*(52), 213–219.

Delgutte, B. (1996). Auditory neural processing of speech. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 507–538). Oxford, England: Blackwell Publishing Ltd.

Delgutte, B., Hammond, B. M., Kalluri, S., Litvak, L. M., & Cariani, P. A. (1996). Neural encoding of temporal envelope and temporal interactions in speech. In W. Ainsworth & S. Greenberg (Eds.), *Proceedings of Auditory Basis of Speech Perception* (pp. 1–9). European Speech Communication Association.

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

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Reviews in Psychology*, *55*, 149–179.

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, *27*(5), 435–443.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, *85*(5), 2154–2164.

Dilley, L., & Pitt, M. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, *21*(11), 1644–1670.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1730–1741.

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

Fowler, C. A., Best, C. T., & McRoberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics*, *48*(6), 559–570.

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

Frazier, J. M., Assgari, A. A., & Stilp, C. E. (2019). Musical instrument categorization is highly sensitive to spectral properties of earlier sounds. *Attention, Perception, & Psychophysics*, *81*(4), 1119–1126.

Fujimura, O., Macchi, M. J., & Streeter, L. A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, *21*(4), 337–346.

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

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517.

Glidden, C. M., & Assmann, P. F. (2004). Effects of visual gender and frequency shifts on vowel category judgments. *Acoustics Research Letters Online*, *5*(4), 132–138.

Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics*, *43*(2), 137–146.

Gottfried, T. L., Miller, J. L., & Payton, P. E. (1990). Effect of speaking rate on the perception of vowels. *Phonetica*, *47*(3–4), 155–172.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, *38*(3), 269–276.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892.

Heffner, C. C., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2013). When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language & Cognitive Processes*, *28*(9), 1275–1302.

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*(3), 964–988.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America*, *120*(5), 2801–2817.

Holt, L. L., & Kluender, K. R. (2000). General auditory processes contribute to perceptual accommodation of coarticulation. *Phonetica*, *57*(2–4), 170–180.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, *108*(2), 710–722.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, *88*(2), 642–654.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384.

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Communication*, *77*, 84–100.

Kidd, G. R., & Wright, B. A. (1994). Improving the detectability of a brief tone in noise using forward and backward masker fringes: Monotic and dichotic presentations. *The Journal of the Acoustical Society of America*, *95*(2), 962–967.

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 736–748.

Kiefte, M., & Kluender, K. R. (2005). Pattern playback revisited: Unvoiced stop consonant perception. *The Journal of the Acoustical Society of America*, *118*(4), 2599–2606.

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. *Journal of the Acoustical Society of America*, *123*(1), 366–376.

Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., & Watsky, S. (2014). Context effects as auditory contrast. *Attention, Perception, & Psychophysics*, *76*, 1437–1464.

Kluender, K. R., & Alexander, J. M. (2007). Perception of speech sounds. In P. Dallos & D. Oertel (Eds.), *The senses: A comprehensive reference* (pp. 829–860). San Diego, CA: Academic.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*(1), 59–69.

Kluender, K. R., & Kiefte, M. (2006). Speech perception within a biologically-realistic information-theoretic framework. In M. A. Gernsbacher & M. Traxler (Eds.), *Handbook of Psycholinguistics* (pp. 153–199). London, England: Elsevier.

Kluender, K. R., Stilp, C. E., & Kiefte, M. (2013). Perception of vowel sounds within a biologically realistic model of efficient coding. In G. Morrison & P. Assmann (Eds.), *Vowel inherent spectral change* (pp. 117–151). Berlin, Germany: Springer.

Kluender, K. R., Stilp, C. E., & Llanos, F. (2019). Longstanding problems in speech perception dissolve within an information-theoretic perspective. *Attention, Perception, & Psychophysics*, *81*(4), 861–883.

Knudsen, V. O. (1929). The hearing of speech in auditoriums. *The Journal of the Acoustical Society of America*, *1*, 56–82.

Ladefoged, P. (1989). A note on "Information conveyed by vowels.". *The Journal of the Acoustical Society of America*, *85*(5), 2223–2224.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lanning, J. M., & Stilp, C. E. (2019). Earlier music biases subsequent musical instrument categorization. *The Journal of the Acoustical Society of America*, *145*, 1822.

Liberman, A. M. (1996). Introduction: Some assumptions about speech and how they changed. In *Speech: A special code*. Cambridge, MA: MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, *65*(4), 497–516.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lindblom, B. E. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 1773–1781.

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, *42*(4), 830–843.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–183.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*(4), 602–619.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *The Journal of the Acoustical Society of America*, *102*(2), 1134–1140.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409.

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

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r". *Cognition*, *24*(3), 169–196.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [∫]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228.

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

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2018). Listening to yourself is special: Evidence from global speech rate tracking. *PLoS One*, *13*(9), 1–19. https://doi.org/10.1371/journal.pone.0203571

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 128–138.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language & Cognitive Processes*, *27*(7–8), 953–978.

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

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 369–378.

Miller, J. L., & Eimas, P. D. (1980). Contextual effects in infant speech perception. *Science*, *209*(4461), 1140–1141.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*(6), 457–465.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512.

Miller, J. L., & Wayland, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, *54*(2), 205–210.

Minifie, F. D., Kuhl, P. K., & Stecher, E. M. (1977). Categorical perception of /b/ and /w/ during changes in rate of utterance. *The Journal of the Acoustical Society of America*, *62*(S1), S79.

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? *Phonetica*, *63*(4), 209–229.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*(1), 365–378.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088–2113.

Nelson, P. C., & Young, E. D. (2010). Neural correlates of context-dependent perceptual enhancement in the inferior colliculus. *The Journal of Neuroscience*, *30*(19), 6577–6587.

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560.

Norris, D., McQueen, M. J., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, 1–17. https://doi.org/10.3389/fpsyg.2012.00320

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pickett, J. M. (1959). Backward masking. *Journal of the Acoustical Society of America*, *31*(12), 1613–1615.

Pickett, J. M., & Decker, L. R. (1960). Time factors in perception of a double consonant. *Language and Speech*, *3*(1), 11–17.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, *34*(4), 314–322.

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics*, *78*(1), 334–345.

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, *7*, 45–56.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, *32*(2), 141–152.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303.

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied PsychoLinguistics*, *37*(6), 1397–1415.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011a). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech*, *54*(2), 147–165.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011b). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 978–996.

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

Repp, B. H. (1982). Phonetic trading relations and context effects - New experimental-evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110.

Repp, B. H. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. *Perception & Psychophysics*, *33*(2), 147–155.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 621–637.

Rysling, A., Jesse, A., & Kingston, J. (2019). Regressive spectral assimilation bias in speech perception. *Attention, Perception, & Psychophysics*, *81*(4), 1127–1146.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception & Psychophysics*, *62*(2), 285–300.

Schouten, J. (1940). The residue and the mechanism of hearing. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, *43*, 991–999.

Shinn, P. C., Blumstein, S. E., & Jongman, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics*, *38*(5), 397–407.

Sjerps, M. J., & Reinisch, E. (2015). Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 710–722.

Sjerps, M. J., & Smiljanic, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics*, *41*(3–4), 145–155.

Sjerps, M. J., Zhang, C., & Peng, G. (2018). Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 914–924.

Srinivasan, N. K., & Zahorik, P. (2013). Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences. *The Journal of the Acoustical Society of America*, *133*(1), EL33–EL39.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *64*(5), 1358–1368.

Stilp, C. E. (2018). Short-term, not long-term, average spectra of preceding sentences bias consonant categorization. *The Journal of the Acoustical Society of America*, *144*(3), 1797.

Stilp, C. E. (2019). Auditory enhancement and spectral contrast effects in speech perception. *Journal of the Acoustical Society of America*.

Stilp, C. E., & Alexander, J. M. (2016). Spectral contrast effects in vowel categorization by listeners with sensorineural hearing loss. *Proceedings of Meetings on Acoustics*, *26*, 1–14. https://doi.org/10.1121/2.0000233

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics*, *72*(2), 470–480.

Stilp, C. E., & Anderson, P. W. (2014). Modest, reliable spectral peaks in preceding sounds influence vowel perception. *Journal of the Acoustical Society of America*, *136*(5), EL383–EL389.

Stilp, C. E., Anderson, P. W., Assgari, A. A., Ellis, G. M., & Zahorik, P. (2016). Speech perception adjusts to stable spectrotemporal properties of the listening environment. *Hearing Research*, *341*, 168–178.

Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. *The Journal of the Acoustical Society of America*, *137*(6), 3466–3476.

Stilp, C. E., & Assgari, A. A. (2017). Consonant categorization exhibits a graded influence of surrounding spectral context. *Journal of the Acoustical Society of America*, *141*(2), EL153–EL158.

Stilp, C. E., & Assgari, A. A. (2019). Natural speech statistics shift phoneme categorization. *Attention, Perception, & Psychophysics*, *81*(6), 2037–2052.

Summerfield, A. Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095.

Summerfield, A. Q., & Assmann, P. F. (1987). Auditory enhancement and speech perception. In M. E. H. Schouten (Ed.), *Proceedings of the NATO advanced research workshop: The psychophysics of speech perception* (pp. 140–150). Dordrecht, The Netherlands: Martinus Nijhoff.

Summerfield, A. Q., & Assmann, P. F. (1989). Auditory enhancement and the perception of concurrent vowels. *Perception & Psychophysics*, *45*(6), 529–536.

Summerfield, A. Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra - phonetic exploration of an auditory aftereffect. *Perception & Psychophysics*, *35*(3), 203–213.

Summerfield, A. Q., Sidwell, A., & Nelson, T. (1987). Auditory enhancement of changes in spectral amplitude. *Journal of the Acoustical Society of America*, *81*(3), 700–708.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301.

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

Utman, J. A. (1998). Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants. *The Journal of the Acoustical Society of America*, *103*(3), 1640–1653.

Viemeister, N. F. (1980). Adaptation of masking. In G. V. D. Brink & F. A. Bilsen (Eds.), *Psychophysical, physiological and behavioural studies in hearing* (pp. 190–198). Delft: Delft University Press.

Viemeister, N. F., & Bacon, S. P. (1982). Forward masking by enhanced components in harmonic complexes. *Journal of the Acoustical Society of America*, *71*(6), 1502–1507.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, *16*(1), 74–79.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1005–1015.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America*, *92*(2), 723–735.

von Békésy, G. (1967). *Sensory perception*. Princeton, NJ: Princeton University Press.

Wade, T., & Holt, L. L. (2005a). Effects of later-occurring nonlinguistic sounds on speech categorization. *The Journal of the Acoustical Society of America*, *118*(3), 1701–1710.

Wade, T., & Holt, L. L. (2005b). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950.

Wark, B., Lundstrom, B. N., & Fairhall, A. (2007). Sensory adaptation. *Current Opinion in Neurobiology*, *17*(4), 423–429.

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, *92*(4), 574–584.

Watkins, A. J. (1988). Spectral transitions and perceptual compensation for effects of transmission channels. In W. Ainsworth & J. Holmes (Eds.), *Proceedings of the 7th symposium of the Federation of Acoustical Societies of Europe: Speech '88*. Edinburgh, England: Institute of Acoustics.

Watkins, A. J. (1989). Spectral transitions and vowel perception. *British Journal of Audiology*, *23*(2), 170.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *90*(6), 2942–2955.

Watkins, A. J. (2005a). Perceptual compensation for effects of echo and of reverberation on speech identification. *Acta Acustica*, *91*(5), 892–901.

Watkins, A. J. (2005b). Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, *118*(1), 249–262.

Watkins, A. J., & Makin, S. J. (1996). Some effects of filtered contexts on the perception of vowels and fricatives. *Journal of the Acoustical Society of America*, *99*(1), 588–594.

Watkins, A. J., & Makin, S. J. (2007). Steady-spectrum contexts and perceptual compensation for reverberation in speech identification. *The Journal of the Acoustical Society of America*, *121*(1), 257–266.

Watkins, A. J., Raimond, A. P., & Makin, S. J. (2011). Temporal-envelope constancy of speech in rooms and the perceptual weighting of frequency bands. *Journal of the Acoustical Society of America*, *130*(5), 2777–2788.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *The Journal of the Acoustical Society of America*, *95*(5), 2694–2701.

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology*, *4*, 1–13.

Winn, M. B., & Stilp, C. E. (2019). Phonetics and the auditory system. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge handbook of phonetics* (pp. 164–192). New York, NY: Routledge.

Zahorik, P., & Brandewie, E. J. (2016). Speech intelligibility in rooms: Effect of prior listening exposure interacts with room acoustics. *The Journal of the Acoustical Society of America*, *140*(1), 74–86.

How to cite this article: Stilp C. Acoustic context effects in speech perception. *WIREs Cogn Sci*. 2019;e1517. https://doi.org/10.1002/wcs.1517