JASA ARTICLE

Short-term, not long-term, average spectra of preceding sentences bias consonant categorization

Anya E. Shorey and Christian E. Stilp[a)]
Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292, USA

ABSTRACT:
Speech sound perception is influenced by the spectral properties of surrounding sounds. For example, listeners perceive /g/ (lower F_3 onset) more often after sounds with prominent high-F_3 frequencies and perceive /d/ (higher F_3 onset) more often after sounds with prominent low-F_3 frequencies. These biases are known as spectral contrast effects (SCEs). Much of this work examined differences between long-term average spectra (LTAS) of preceding sounds and target speech sounds. *Post hoc* analyses by Stilp and Assgari [(**2021**) Atten. Percept. Psychophys. 83(6) 2694–2708] revealed that spectra of the last 475 ms of precursor sentences, not the entire LTAS, best predicted biases in consonant categorization. Here, the influences of proximal (last 500 ms) versus distal (before the last 500 ms) portions of precursor sentences on subsequent consonant categorization were compared. Sentences emphasized different frequency regions in each temporal window (e.g., distal low-F_3 emphasis, proximal high-F_3 emphasis, and vice versa) naturally or *via* filtering. In both cases, shifts in consonant categorization were produced in accordance with spectral properties of the proximal window. This was replicated when the distal window did not emphasize either frequency region, but the proximal window did. Results endorse closer consideration of patterns of spectral energy over time in preceding sounds, not just their LTAS. © *2023 Acoustical Society of America*.
https://doi.org/10.1121/10.0017862
(Received 18 November 2022; revised 31 March 2023; accepted 31 March 2023; published online 20 April 2023)
[Editor: Sven Mattys] Pages: 2426–2435

I. INTRODUCTION

All perception takes place in context. This is particularly true when perceiving speech, as recognition of a given speech sound is heavily influenced by the acoustic properties of surrounding sounds (Ainsworth, 1975; Nearey, 1989). For example, perception of crucial spectral properties in a sound is often relative to spectral properties of surrounding sounds (referred to as "spectral context effects"); likewise, perception of key temporal properties in a given sound is often relative to the temporal properties of surrounding sounds ("temporal context effects"). These acoustic context effects are pervasive in speech perception (see Stilp, 2020, for review).

Acoustic context offers resilience against the extreme acoustic variability of the speech signal. While absolute acoustic properties are highly variable from moment to moment and from sound to sound, perception can proceed by operating on relative comparisons (i.e., how the acoustic properties of a target sound compare to those of the surrounding context). These context effects can occur on proximal (e.g., the sound immediately preceding the target sound) or distal time scales (e.g., further displaced in time from, and nonadjacent to, the target sound). Yet, the extreme variability in speech necessitates that the context itself is an evolving basis of comparison. This evolution might challenge how these contexts on different time scales relate to one another. The question at hand in the present investigation is not whether acoustic context effects can influence speech perception; instead, the question is *when* these effects influence perception. Global acoustic characteristics of contexts tend to be fairly stable across time (e.g., owing to speaking at a particular rate or to resonances occurring at relatively higher or lower frequencies as a function of vocal tract length), but these characteristics are free to abruptly change on a shorter time scale (e.g., owing to temporal and/or spectral properties of a few syllables or words being spoken). For example, when distal context is influencing perception in a particular direction (e.g., slow speaking rate promoting perception of subsequent fast rate) but proximal context is influencing perception in the opposite direction (e.g., fast speaking rate promoting perception of subsequent slow rate), which contextual influence (if any) informs perception of the subsequent target?

Studies of temporal context effects (also termed "speaking rate normalization") are relatively uniform in their answer to this question: proximal context exerts a much larger influence on perception than distal context. This was first reported by Summerfield (1981), in which speaking rate for the final word in the context phrase "why are you" was much more effective than that for the first two context words in shifting voicing perception in the target words /biz/-/piz/ ("bees"-"peas"). Kidd (1989) reported similar findings for the context phrase "A bird in the hand is worth two in the" influencing perception of voicing in the target words /gi/-/ki/ ("gee"-"kee"). Even when the rest of

[a)]Electronic mail: christian.stilp@louisville.edu

 0001-4966/2023/153(4)/2426/10/$30.00

https://doi.org/10.1121/10.0017862

the context phrase was spoken quickly (or slowly), a slow (or fast) rate for the final context word was sufficient to determine the direction of the shift in perception of the target. Reinisch *et al.* (2011) varied context sentence rates to measure perception of the presence of certain speech sounds (/s/, /t/) at the beginning of the target words (fast context sentences promoted perception of longer-duration context words with these speech sounds present). When distal and proximal speaking rates were in conflict, the proximal rate exerted a much larger influence on target word perception than the distal rate. Similar observations have been reported when participants were tasked with detecting the presence or absence of a function word (Heffner *et al.*, 2013). Finally, on a longer time scale, Reinisch (2016) first exposed listeners to an extended conversation where two talkers spoke at slow and fast rates, respectively (the distal context), then presented trials in which a context sentence spoken at different rates (the proximal context) preceded the target vowel. Again, temporal properties of the proximal context exerted a much greater influence on the perception of the target words. While it bears noting that a competing distal context might modestly attenuate the effect produced by the proximal context (e.g., slowly spoken distal context attenuating the context effect produced by the quickly spoken proximal context), across stimuli and paradigms, proximal context appears to exert a much stronger influence on perception of temporal properties in the target sound/word than the distal context does.

The relative influences of proximal and distal contexts for spectral contrast effects (SCEs) have been far less studied and with less clear results. Holt (2006) presented a 2100-ms series of pure tones before a /dɑ/-/gɑ/ target syllable on each trial. The tone series was divided into three successive 700-ms epochs, each with a different mean frequency (1800, 2300, or 2800 Hz). These mean frequencies were tested in all possible orders. Lower frequencies in the last 700-ms epoch (which immediately preceded the target consonant) were predicted to promote more /dɑ/ (higher F_3 onset) percepts, and higher frequencies in the last epoch were predicted to promote more /gɑ/ (lower F_3 onset) percepts. Statistical analyses failed to find consistent effects of any of these context sequences on /dɑ/-/gɑ/ categorization, leading Holt (2006) to conclude that proximal spectral context was far less perceptually salient than the global statistical structure (the mean, or long-term average spectrum, LTAS, across all windows). Later, Stilp and Assgari (2021) measured how context sentences with different spectral compositions influenced perception of /dɑ/-/gɑ/ targets. Rather than filtering the sentences, they selected and presented sentences whose spectra inherently possessed the desired properties (as part of a "natural signal statistics" approach to studying context effects; see also Stilp and Assgari, 2019). A total of 12 SCEs were measured across five different experiments. Unlike Holt (2006), the LTAS of context sentences were poor predictors of context effect magnitudes ($r = 0.21$); instead, spectra of the last 475 ms of context sentences (i.e., proximal context) were excellent predictors ($r = 0.90$). However, the reverse correlation analyses that uncovered this result were *post hoc*; sentences were initially chosen as stimuli based on their long-term spectral properties and not those near sentence offset. Given this, distal spectral properties of sentences (everything preceding the last 475 ms) were uncontrolled and highly variable across items. Thus, it remains unclear whether SCEs pattern similarly to temporal context effects; specifically, whether proximal spectral context influences subsequent speech perception more than distal context does.

The present experiment tested how competing spectral contexts influenced perception of speech sounds (/dɑ/-/gɑ/). Context sentences were selected and presented based on specific patterns of spectral properties over time. One pattern tested was sentence contexts with a strongly biased distal window (e.g., stronger spectral energy in the low-F_3 frequency region) and a strongly and oppositely biased proximal window (e.g., stronger spectral energy in the high-F_3 frequency region, or vice versa). Another context pattern tested was relatively neutral distal window (approximately equal spectral energy in low-F_3 and high-F_3 frequency regions) followed by a strongly biased proximal window (stronger spectral energy in either the low-F_3 or high-F_3 frequency region). While Stilp and Assgari (2021) observed a strong influence of the last 475 ms of context sentences on perception of /dɑ/-/gɑ/ targets *post hoc*, the present study is a direct test of the sufficiency of contextual spectral properties in this temporal window. Consistent with the literature on temporal context effects reviewed previously, we predicted that the proximal context would exert a stronger influence on speech sound categorization than the distal context. Additionally, consistent with the patterns of results reported by Stilp and Assgari (2019, 2021), SCEs produced by filtered context sentences were predicted to be larger than SCEs produced by unfiltered context sentences.

II. METHODS

A. Participants

Fifty-one undergraduate students at the University of Louisville participated in exchange for course credit. All reported being native English speakers with no known hearing impairments. This study was approved by the Institutional Review Board of the University of Louisville, and all participants provided informed consent at the beginning of the experiment. Two additional participants began the experiment but did not complete the test blocks, so their responses were removed before data analysis.

B. STIMULI

1. Mean square difference (MSD) calculation

Central to the selection and creation of all context sentence stimuli was the calculation of MSDs. These calculations followed the same protocol as detailed in Stilp and Assgari (2021). Briefly, each sentence was analyzed using two separate bandpass filters, with the passband at either 1700–2700 Hz or 2700–3700 Hz. Transition regions

between the passband and stopbands were 5 Hz. Filters were created using the fir2 command in MATLAB (The MathWorks Inc., 2021) using 1000 coefficients. Each resulting speech band was then rectified and low-pass filtered (2nd-order Butterworth filter with 30-Hz cutoff) to obtain its amplitude envelope. The root mean square (RMS) energy for each envelope was then converted into dB. The MSD was defined as the difference in energy across these two frequency regions (always subtracted as low-F_3 energy minus high-F_3 energy, with positive MSDs indicating more energy in the low-F_3 region and negative MSDs indicating more energy in the high-F_3 region). Critically, for each sentence, MSDs were calculated for two distinct temporal intervals: the last 500 ms of the sentence (the Proximal window, which immediately preceded the target stimulus on each trial) and the rest of the sentence from its onset up to the last 500 ms (the Distal window, which was temporally nonadjacent to the target stimulus on each trial).

2. Unfiltered contexts

To identify candidate stimuli, MSDs in the Distal and Proximal windows were calculated on all items in two large corpora of sentences spoken in American English: the TIMIT database (630 talkers each speaking ten sentences; Garofolo *et al.*, 1990) and the HINT database (one talker speaking 275 unique sentences; Nilsson *et al.*, 1994). Across databases, the largest differences between Distal and Proximal MSDs (which were of interest for creating the Distal Competing condition described in the following) were comparable. However, the HINT database guaranteed that all of these items were spoken by the same talker whereas the TIMIT database did not. Additionally, in an analysis of the entire TIMIT corpus, sentences generally possessed positive MSDs, or more energy in the low-F_3 frequency region (1700–2700 Hz) than the high-F_3 frequency region (2700–3700 Hz; see Fig. 2 from Stilp and Assgari, 2021). Thus, while the TIMIT corpus provides an abundance of options for candidate sentences with positive MSDs, it provides far fewer options for candidate sentences with negative MSDs. Conversely, analysis of the entire HINT corpus revealed a more balanced distribution of MSDs, with several candidate sentences that had positive MSDs as well as candidate sentences with negative MSDs. Therefore, four sentences were selected from the HINT database as stimuli here.

Sentences presented in the unfiltered conditions met three criteria. First, MSDs in the Distal and Proximal windows of these sentences naturally followed the desired patterns for the present experiment. For instance, each of the four sentences exhibited a strong bias in the Proximal window (a strong positive MSD indicating greater energy in low-F_3 frequencies, or a strong negative MSD indicating greater energy in high-F_3 frequencies); further details of these spectral properties are provided in the following. Second, all four sentences were spoken by the same talker, which constrained the extent of acoustic variability across items. Large acoustic variability across various context sentences spoken by different talkers (e.g., pitch, speaking rate, duration, etc.) is thought to diminish SCE magnitudes (Stilp and Assgari, 2019, 2021). Sentences in the HINT database are generally consistent in their speaking rates, semantic complexity, and syntactic construction, all of which lowers item-to-item acoustic variability. Third, the talker was an adult man, which was also the case for the filtered context sentence and the target syllables detailed in the following. While the talkers who produced the unfiltered context sentences, the filtered context sentence, and the target syllables all differed, it bears noting that matching the talker across context and target stimuli is not a prerequisite for producing SCEs. These effects have been reported when the speech targets were preceded by speech contexts spoken by a different talker (Watkins, 1991; Lotto and Kluender, 1998) or by nonspeech contexts (Lotto and Kluender, 1998; Holt, 2006; Stilp, 2020).

Four unfiltered sentences were presented as context stimuli: (1) high-F_3-emphasized-Distal with low-F_3-emphasized-Proximal, (2) low-F_3-emphasized-Distal with high-F_3-emphasized-Proximal, (3) neutral-Distal with low-F_3-emphasized-Proximal, (4) neutral-Distal with high-F_3-emphasized-Proximal. Each sentence is illustrated in Fig. 1. Sentences with high-F_3-emphasized-Distal/low-F_3-emphasized-Proximal [Fig. 1(A)] and low-F_3-emphasized-Distal / high-F_3-emphasized-Proximal [Fig. 1(B)] were paired together to form the Unfiltered Distal Competing condition, where one of these two sentences was presented on each trial. Sentences with neutral-Distal/low-F_3-emphasized-Proximal [Fig. 1(C)] and neutral-Distal / high-F_3-emphasized-Proximal [Fig. 1(D)] were paired together to form the Unfiltered Distal Neutral condition; again, one of these two sentences was presented on each trial. Details of each sentence follow in the next paragraph.

In the Distal Competing condition, MSDs were biased toward one frequency region in the Distal window of the sentence and then biased toward the other frequency region in the Proximal window. Figure 1(A) depicts the context sentence "She looked in her mirror" (duration = 1532 ms). The Distal window had a large negative MSD (–11.32) particularly due to the higher-frequency frication energy of /ʃ/ in "she" and F_3 and F_4 in the vowel /ʊ/ in "look." The Proximal window had a large positive MSD (9.80) owing to lower-frequency energy in the frequencies of F_2 and F_3 throughout the word "mirror." Figure 1(B) depicts the context sentence "The family bought a house" (duration = 1586 ms). The Distal window had a large positive MSD (10.43) owing to greater lower-frequency energy at F_2 and F_3 during /ae/ of "family" and the interval combining the release of /t/ in "bought," the /ʌ/ of "a," and /h/ in "house." The Proximal window had a large negative MSD (–8.57) owing to the higher-frequency frication energy in /s/ of "house."

In the Distal Neutral condition, MSDs were near-zero during the Distal window of the sentence before being strongly biased toward one frequency region in the Proximal window. Figure 1(C) depicts the context sentence "Father forgot the bread" (duration = 1601 ms). The Distal window exhibited comparable energy across low-F_3 and high-F_3

https://doi.org/10.1121/10.0017862

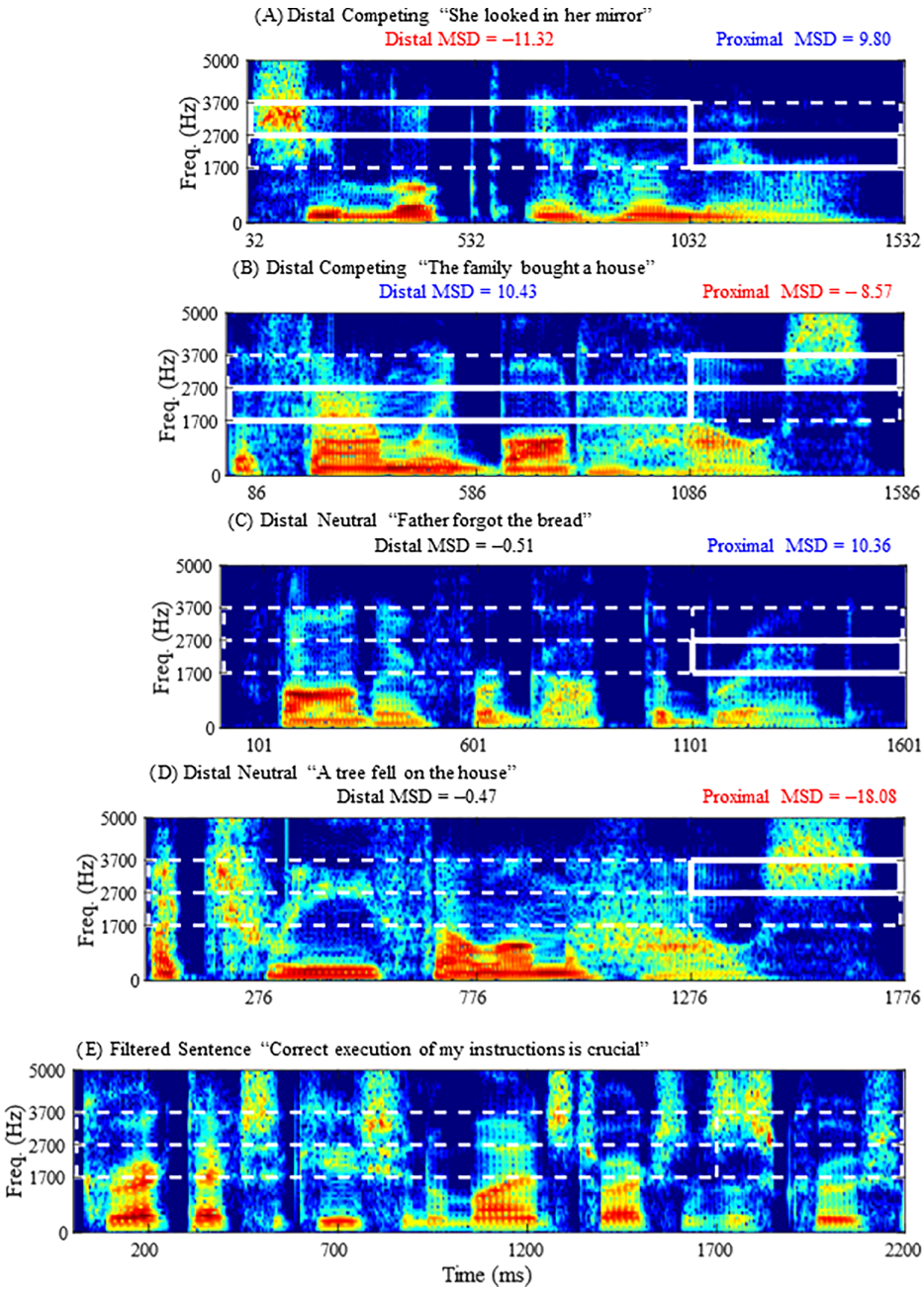

FIG. 1. (Color online) Annotated spectrograms of the context sentences aligned at their offsets. White lines denote low-F_3 (1700–2700 Hz) and high-F_3 (2700–3700 Hz) frequency regions in the proximal windows (last 500 ms of each sentence) and the distal windows (from onset until 500 ms from sentence offset) of each sentence. MSDs are noted in the figure titles to indicate context sentence windows (distal or proximal) that have a relative emphasis on the low-F_3 frequency region or relative emphasis on the high-F_3 frequency region; regions of emphasis are denoted by solid white lines. MSD values that markedly differ from 0, and therefore are predicted to influence speech target categorization, are colored blue (more energy in low-F_3 frequencies) or red (more energy in high-F_3 frequencies); near-zero MSDs are written using black text. Figure 1(E) illustrates the sentence presented in filtered conditions, but this token has not been filtered to emphasize either frequency region.

regions (MSD = –0.51) before becoming strongly biased toward the low-F_3 region in the Proximal window (MSD = 10.36) owing to the frequencies of F_2 and F_3 throughout the word "bread." Figure 1(D) depicts the context sentence "A tree fell on the house" (duration = 1776 ms). The Distal window again exhibited comparable energy across low-F_3 and high-F_3 regions (MSD = –0.47) before becoming strongly biased toward the high-F_3 region in the Proximal window (MSD = –18.08) due to the frication energy of /s/ in "house."

3. Filtered contexts

A single sentence was selected from the TIMIT database (Garofolo *et al.*, 1990): an adult man saying, "Correct execution of my instructions is crucial" [Fig. 1(E); 2200 ms]. Filtered renditions of this stimulus have been highly successful in biasing consonant categorization in previous studies (Stilp and Assgari, 2017, 2021). Here, this sentence was presented in two experimental conditions: Filtered Distal Competing and Filtered Distal Neutral. In each of these conditions, this sentence was presented on every trial. The only trial-to-trial variability in the context sentence was its spectral properties, which were engineered to match spectral properties of the unfiltered sentences, as described in the following.

First, this sentence was divided into its Distal (first 1700 ms) and Proximal (last 500 ms) segments. The native MSD of the Distal segment was 1.85, indicating relatively equal energy across low-F_3 and high-F_3 regions. The native

MSD of the Proximal segment was −10.82 owing to the higher-frequency frication noise in /z/ of "is" and /ʃ/ in the middle of "crucial." Critically, MSDs were changed via filtering in order to match the MSDs in the unfiltered sentences described previously. Filtering followed the same procedures outlined in Stilp and Assgari (2021), with 1000-Hz-wide finite impulse response filters (again spanning 1700–2700 Hz or 2700–3700 Hz) generated in MATLAB using the fir2 function with 1200 coefficients. For example, after excising the Distal segment of this sentence, one second of silence was prepended and appended to the segment to control for filter delay. To match the MSD of the Distal window in one of the unfiltered sentences (e.g., in "She looked in her mirror," Distal MSD = –11.32), filter gain for amplifying the relevant frequency region (here, 2700–3700 Hz, in order to make the MSD more strongly negative) was adjusted iteratively until its MSD was within 0.1 dB of the target MSD in the unfiltered sentence. This process was repeated for the Proximal segment of this sentence, excising it and prepending and appending one second of silence to it. To match the MSD of the Proximal window in the same unfiltered sentences (e.g., in "She looked in her mirror," Proximal MSD = 9.80), filter gain for amplifying the relevant frequency region (here, 1700–2700 Hz, in order to make the MSD positive) was adjusted iteratively until its MSD was within 0.1 dB of the target MSD in the unfiltered sentence. Finally, these Distal and Proximal segments of the filtered sentence were concatenated, resulting in a context sentence whose Distal and Proximal MSDs matched those of a corresponding unfiltered sentence (in this example, one of the Distal Competing stimuli). This process was repeated so that a filtered sentence shared the same Distal MSD (10.43) and Proximal MSD (−8.57) as those in the other unfiltered sentence in that condition ("The family bought a house" in the Distal Competing condition).

In all, four filtered context sentence stimuli were created. In the Filtered Distal Competing condition, one stimulus shared MSDs with "She looked in her mirror" (Distal MSD = –11.32, Proximal MSD = 9.80), and another stimulus shared MSDs with "The family bought a house" (Distal MSD = 10.43, Proximal MSD = –8.57). Therefore, listeners heard the same patterns of MSDs in the Filtered Distal Competing block as in the Unfiltered Distal Competing block. In the Filtered Distal Neutral condition, one stimulus shared MSDs with "Father forgot the bread" (Distal MSD = –0.51, Proximal MSD = 10.36), and another stimulus shared MSDs with "A tree fell on the house" (Distal MSD = –0.47, Proximal MSD = –18.08). Therefore, on each trial, listeners heard the same patterns of MSDs in the Filtered Distal Neutral block as in the Unfiltered Distal Neutral block.

4. Targets

Target consonants were a series of ten morphed natural tokens from a continuum ranging from a /gɑ/ endpoint to a /dɑ/ endpoint (Stephens and Holt, 2011). These syllables were the same targets as tested in previous studies of SCEs in consonant categorization (Stilp and Assgari, 2017, 2021). F_3 onset frequencies in these resynthesized real speech tokens varied from 2338 Hz (/gɑ/ endpoint) to 2703 Hz (/dɑ/ endpoint) before converging at/near 2614 Hz for the following /ɑ/. The duration of the consonant transition was 63 ms, and the total syllable duration was 365 ms.

All context sentences and target syllables were low-pass filtered at 5 kHz and set to equal RMS amplitude. Experimental trials were then created by concatenating each target syllable to each context sentence with 50-ms silent interstimulus intervals. All experimental trials are available at https://osf.io/95z42/.

C. Procedure

Participants were seated in a sound attenuating booth (Acoustic Systems, Inc., Austin, TX). Stimuli were D/A converted by RME HDSPe AIO sound cards (Audio AG, Haimhausen, Germany) on personal computers and passed through a programmable attenuator (TDT PA4, Tucker-Davis Technologies, Alachua, FL) and headphone buffer (TDT HB6). Stimuli were presented diotically at an average of 70 dB sound pressure level (SPL) over circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY). A custom MATLAB script led the participants through the experiment.

Participants first completed 20 practice trials. On each practice trial, the context was a unique sentence from the AzBio corpus (Spahr *et al.*, 2012). This prevented listeners from becoming overly familiar with any of the talkers who produced context sentences in the main experiment. Ten sentences were filtered to amplify the low-F_3 (1700–2700 Hz) frequency region by 20 dB; ten other sentences were filtered to amplify the high-F_3 (2700–3700 Hz) frequency region by 20 dB. This filtering was done to expose listeners to some of the more extreme MSDs they would encounter in the main experiment. On each trial, the target was either the /gɑ/ or the /dɑ/ endpoint from the consonant continuum. After each trial, participants clicked the mouse to indicate whether the target syllable sounded more like the sound "ga" or "da." Listeners were required to categorize consonants with at least 80% accuracy within three attempts of the practice block to proceed to the main experiment.

The main experiment comprised four blocks of 160 trials apiece (2 context sentences x 10 targets x 8 repetitions). In the Unfiltered Distal Competing block, each trial presented either "She looked in her mirror" [low-F_3-biased Proximal window; Fig. 1(A)] or "The family bought a house" [high-F_3-biased Proximal window; Fig. 1(B)] before the target syllable. In the Unfiltered Distal Neutral block, each trial presented either "Father forgot the bread" [low-F_3-biased Proximal window; Fig. 1(C)] or "A tree fell on the house" (high-F_3-biased Proximal window; Figure 1D) before the target syllable. In Filtered Distal Competing and Filtered Distal Neutral blocks, each trial presented the sentence "Correct execution of my instructions is crucial" [Fig. 1(E)]. This token was filtered to match the MSDs of either

https://doi.org/10.1121/10.0017862

the Unfiltered Distal Competing sentences (thus creating the Filtered Distal Competing block) or the Unfiltered Distal Neutral sentences (thus creating the Filtered Distal Neutral block). These four blocks were presented in counterbalanced orders across participants, and trials within each block were randomized. No feedback was provided. The experiment was self-paced and participants had the opportunity to take short breaks between each block as needed. The total experimental session lasted approximately one hour.

III. RESULTS

A performance criterion was implemented such that participants were required to maintain 80% accuracy on continuum endpoints throughout the main experiment. One participant failed to meet this criterion and so was excluded from data analyses; the final sample size consisted of 50 participants. Logistic regressions were fit to each participant's responses at each level of Spectral Peak (low-F_3 and high-F_3 emphasis in the proximal window) for each block (Filtered Distal Neutral, Filtered Distal Competing, Unfiltered Distal Neutral, Unfiltered Distal Competing). Then, SCEs were calculated as the number of continuum steps separating 50% points on psychometric functions (i.e., 50% on the Proximal-Low-F_3 function and 50% on the Proximal-High-F_3 function). The grand-mean SCE was 0.740 steps [standard error of the mean (SEM) = 0.089] for Filtered Distal Neutral, 0.209 steps (SEM = 0.088) for Filtered Distal Competing, 0.528 steps (SEM = 0.087) for Unfiltered Distal Neutral, and 0.238 steps (SEM = 0.066) for Unfiltered Distal Competing. Individuals' SCE magnitudes and group mean SCE magnitudes and standard errors are depicted in Fig. 2(b).

Results were also analyzed at the group level. The dependent variable is the listener's response, which is a binary outcome variable for which a response of "ga" = 0 and "da" = 1. Therefore, data were analyzed with a generalized linear mixed effects model in R (R Core Team, 2021) using the lme4 package (Bates *et al.*, 2015) with glmerControl and the bobyqa optimizer added. The model contained fixed effects of theoretical interest including the main effect of Target (mean-centered), the main effect of Spectral Peak (sum-coded, High F_3 coded as –0.5 and Low F_3 coded as +0.5), the main effect of Condition (sum-coded, Filtered coded as –0.5 and Unfiltered coded as +0.5), and the main effect of Timing (sum-coded, Distal Competing coded as –0.5 and Distal Neutral coded as +0.5), and the interactions between Spectral Peak and Condition, and Spectral Peak and Timing. The random slopes were built iteratively, adding one slope at a time until all main effects were included as slopes, or the model failed to converge/explain significantly more variance. The final model included random slopes by Target, Spectral Peak, Condition, and Timing, and random intercepts by participant. All data and annotated results scripts are available at https://osf.io/95z42/.

Model estimates and significance values are presented in Table I and are plotted in Fig. 2(a). The primary effect of interest was that of Spectral Peak (which relates to the magnitude of spectral contrast effects) and its interactions. Spectral Peak was a significant predictor, such that stimuli that emphasized low frequency-content in the Proximal window (low F_3) were likely to yield more "da" responses (main effect of Spectral Peak), thus confirming the presence of SCEs. The significant Spectral Peak by Timing interaction indicates that the magnitudes of SCEs were larger in the Distal Neutral than in the Distal Competing sentences. The magnitudes of SCEs were not different in the Filtered Conditions and Unfiltered Conditions (n.s. Spectral Peak by condition interaction).

Other significant results include that the log odds of responding "da" was higher than the likelihood of responding "ga" overall (significant intercept). Target was a significant predictor of responses, such that more /dɑ/-like target stimuli prompted a higher likelihood of "da" responses (main effect of Target). Finally, the log odds of responding "da" were higher for the Unfiltered sentences than the Filtered sentences (main effect of condition).

Group-level SCEs were calculated using the same generalized linear mixed-effects model posted previously but with different coding schemes for the fixed effects. Categorical coding was utilized to set one level of condition (Filtered or Unfiltered) and one level of Timing (Distal Competing or Distal Neutral) as the default condition (e.g., Filtered Distal Competing as the default). In doing so, the fixed effect of Spectral Peak tested whether the SCE in a given block (again measured as the number of continuum steps separating 50% points on psychometric functions) significantly differed from zero. All SCEs significantly differed from zero (Filtered Distal Neutral: SCE = 0.657 steps, $Z = 10.258$, $p < 0.001$; Filtered Distal Competing: SCE = 0.282 steps, $Z = 4.434$, $p < 0.001$; Unfiltered Distal Neutral: SCE = 0.585 steps; $Z = 9.147$, $p < 0.001$; Unfiltered Distal Competing: SCE = 0.210 steps, $Z = 3.304$, $p = 0.001$).

The powerSim function of the simr package in R (Green and MacLeod, 2016) was utilized to calculate the observed power for the interactions in the mixed-effects model. Based on 40 simulations, the statistically significant interaction between Spectral Peak × Timing possessed 100% power (effect size = 0.60), but the non-significant interaction between Spectral Peak × Condition possessed only 25% power. These power analyses confirm that the Spectral Peak × Timing interaction was adequately powered at this sample size; conversely, the Spectral Peak × Condition interaction appears to be more likely a true null result at this sample size rather than underpowered.

IV. DISCUSSION

Speech sound recognition is heavily influenced by preceding acoustic context; however, disagreement exists as to whether the proximal or distal context more heavily influences perception of later sounds. Using pure tone contexts with different mean frequencies, Holt (2006) reported no

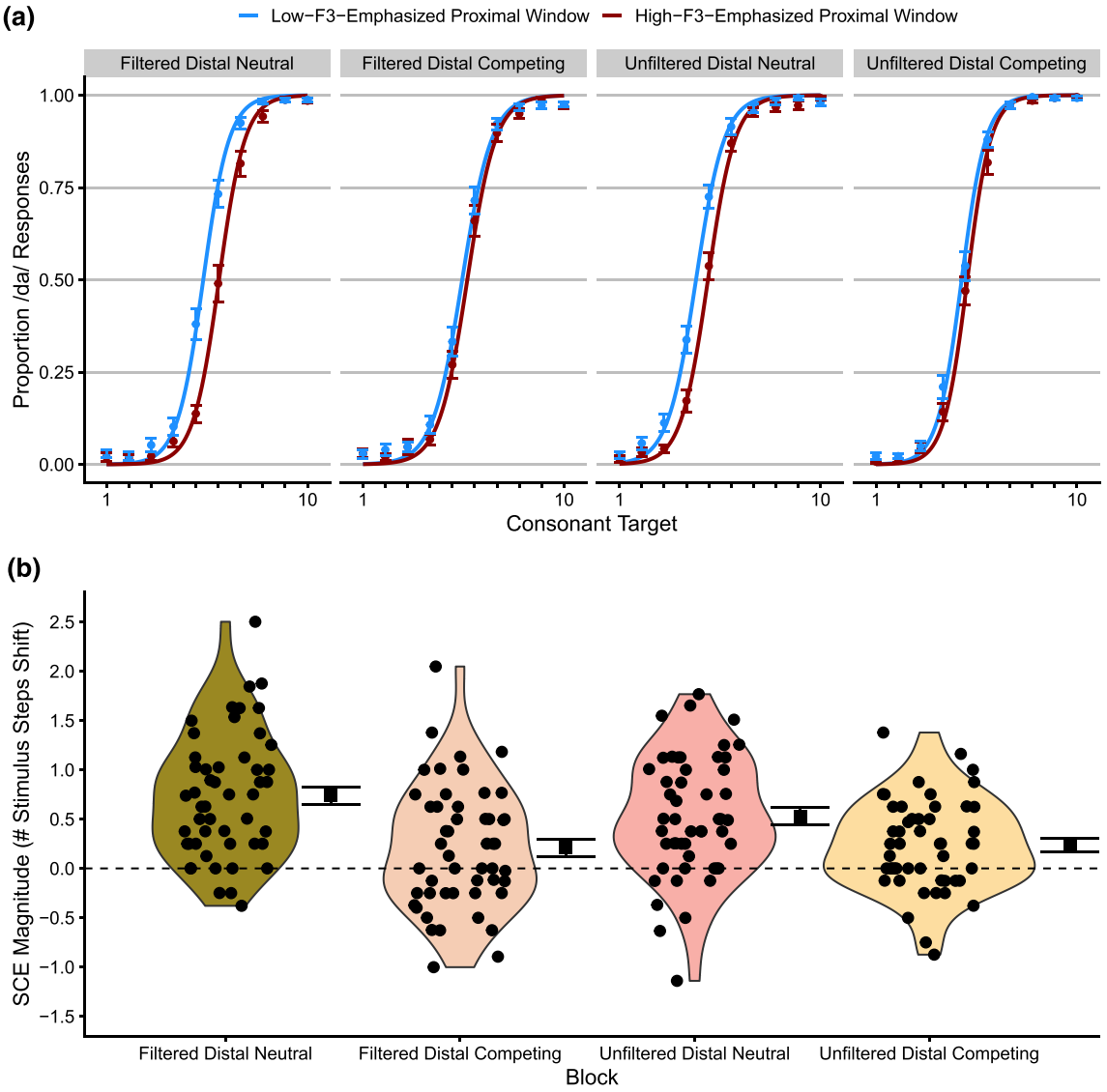

FIG. 2. (Color online) (a) Group level model results with raw means and ± one standard error of the mean overlayed. The panels depict the proportion of "da" responses as a function of consonant target from the ten-step consonant continuum. For plotting purposes only, the model had fixed effects of the interaction of Target, Spectral Peak, and Block (Filtered Distal Neutral, Filtered Distal Competing, Unfiltered Distal Neutral, Unfiltered Distal Competing), and their main effects (the random effect structure was the same as the model reported in the main text). Lighter (blue) lines represent conditions in which low F_3 frequencies are emphasized in the proximal window and darker (red) lines represent conditions in which high F_3 frequencies are emphasized in the proximal window. (b) Individual participants' spectral contrast effect magnitudes by condition. Each circle represents a participant within a block. Squares represent group means and error bars represent ± one standard error.

TABLE I. Beta estimates ($\hat{\beta}$), *SE*, *Z*, and *p* values for the fixed effects of the mixed effects model. As described in the main text, Target was entered in the model as a continuous factor, centered around the mean. Spectral Peak, Condition, and Timing were sum-coded; the level associated with the −0.5 contrast for each factor is shown in parentheses.

Predictors	$\hat{\beta}$	*SE*	*Z*	*p*
Intercept	0.457	0.112	4.097	<0.001
Target	1.650	0.069	23.979	<0.001
Spectral Peak (High F_3)	0.716	0.086	8.341	<0.001
Condition (Filtered)	1.298	0.159	8.187	<0.001
Timing (Distal Competing)	0.114	0.081	1.407	0.159
Spectral Peak × Condition	−0.119	0.086	−1.378	0.168
Spectral Peak × Timing	0.619	0.087	7.141	<0.001

systematic influence of local spectral detail on the perception of /dɑ/-/gɑ/ target sounds, instead suggesting the long-term average spectrum of the context was its most perceptually salient characteristic. Later, Stilp and Assgari (2021) reported that spectral properties of the last 475 ms of context sentences exerted the strongest influence on target /dɑ/-/gɑ/ perception, but their analyses of sentence spectra were conducted *post hoc*. Here, context sentences were selected or constructed to have specific patterns of spectral properties over time: the last 500 ms (the proximal context) was predicted to bias /dɑ/-/gɑ/ perception in one direction, and the spectrum of the rest of the sentence (everything preceding those last 500 ms; the distal context) was either neutral or in direct competition with that influence. In all

https://doi.org/10.1121/10.0017862

experimental conditions tested, the proximal context biased target perception.

Speech sound categorization being more sensitive to proximal spectral context than distal spectral context is consistent with similar studies conducted in the temporal domain (i.e., speaking rate normalization, or temporal contrast effects). Across a wide range of stimuli and experimental paradigms, the temporal characteristics of proximal context exerted a much stronger influence on perception of temporal properties in the target sound/word than distal context did (Summerfield, 1981; Kidd, 1989; Reinisch *et al.*, 2011; Heffner *et al.*, 2013; Reinisch, 2016). These findings suggest that while long-term characteristics can be important for perception, it is equally if not more important to remain sensitive to local deviations. This pattern of results is in keeping with an evolutionary advantage to maintaining awareness to the current environment. Stable environmental properties may well inform adaptive behavior, but events and properties of the most recent past may prove most informative and consequential for perception and action. More generally, this is consistent with the notion of perception as operating as a change detector; changes in local statistics and properties are often accentuated by the auditory system (von Békésy, 1967; Warren, 1985; Kluender *et al.*, 2003; Winn and Stilp, 2019; Stilp, 2020).

There are two primary differences between the methodologies of Holt (2006), Stilp and Assgari (2021), and the present study. First, there can be myriad acoustic differences across speech and nonspeech stimuli (such as the pure tones used in Holt, 2006). Not all nonspeech sounds model acoustic characteristics of speech equally well (Stilp *et al.*, 2022). These differences are most evident in the spectral complexity of the pure tones deployed by Holt (2006), which was the primary contributing factor for SCEs. While using pure tones offers great experimental control, future studies interested in speech perception may be well served by using sounds with greater spectrotemporal complexity (more on par with that of speech) to increase ecological validity. The comparison with an unfiltered condition moves this line of inquiry closer to natural listening conditions by using naturally produced speech contexts and speech targets (Stilp and Assgari, 2019, 2021). Alternatively, if the study is using nonspeech stimuli in an effort to isolate and study one (acoustic) aspect of speech, justification as to which aspects are deliberately being modeled by the nonspeech and which are not should be explicitly discussed (Stilp *et al.*, 2022).

Second, in the present study, the total duration of the context sentences was of variable length, but the time course of the proximal window was always held constant. A duration of 500 ms was selected for the proximal window since Stilp and Assgari (2021) found that spectral characteristics of the last 475 ms of a sentence were the most highly correlated with spectral contrast effect magnitudes. Alternatively, Holt (2006) defined the proximal window as the last 700 ms of the preceding context. Calculating the MSD of the last 700 ms of the present stimuli (instead of the last 500 ms) has variable impacts. Sentences with prominent low-F_3 frequencies in the proximal window saw their MSDs decrease slightly (by $\approx$2 dB), but sentences with prominent high-F_3 frequencies in the proximal windows saw their MSDs decrease dramatically (by $\approx$10 dB). However, this *post hoc* analysis is only so informative given that candidate sentences were analyzed and stimuli were selected specifically due to MSDs in the last 500 ms (700-ms proximal window durations were not considered at the time). In the reverse correlation analyses by Stilp and Assgari (2021), the optimal predictor of SCE magnitudes was a proximal window duration of 475 ms ($r = 0.90$), but proximal window durations of 500 ms ($r = 0.86$) and 700 ms ($r = 0.75$) were still both effective predictors. Therefore, the duration of the proximal window, be it 500 or 700 ms, does not appear to be the reason the results of Holt (2006) and the present study diverge. There is no functional significance ascribed to a proximal window duration of 500 ms, 700 ms, or any other specific value; at present, there is no standard definition of what constitutes the proximal or distal window of a context (see Heffner *et al.*, 2017 for further discussion).

Further differences between studies are revealed when comparing local versus global time scales of the acoustic contexts. Holt (2006) reported similar categorization behavior across all local contexts (spectral properties in the last 700 ms preceding the target), deducing that listeners were instead utilizing global contextual information (the LTAS of the entire 2100-ms context sequence, which was held constant across conditions). While Stilp and Assgari (2021) did not select or construct stimuli according to their local (proximal) spectral characteristics, they did observe that these proximal characteristics were excellent predictors of SCE magnitudes and that global characteristics (MSDs calculated across the full duration of context sentences) were very poor predictors. Here, stimuli were selected or constructed to compare the proximal window (last 500 ms) and the distal window (everything preceding the last 500 ms; this varied by token) irrespective of their global spectral characteristics. However, the global characteristics of these stimuli make markedly different predictions than the local (i.e., proximal) characteristics do. Full-sentence MSDs for sentences in the Distal Competing condition actually predict perceptual shifts in the *opposite* direction of what occurred. "She looked in her mirror" has a full-sentence MSD of −10.71 (more closely in line with the Distal MSD of –11.32 than the Proximal MSD of 9.80); "The family bought a house" has a full-sentence MSD of 8.37 (again more closely in line with the Distal MSD of 10.43 than the Proximal MSD of –8.57). These longer-term spectral characteristics would predict that "She looked in her mirror" would produce more lower-frequency "ga" responses and that "The family bought a house" would produce more higher-frequency "da" responses, but the shift was in the opposite direction (and in the direction predicted by the Proximal MSDs; Fig. 1). In the Distal Neutral condition, full-sentence MSDs were near-zero ("Father forgot the bread" full-sentence MSD = 0.59; "A tree fell on the house" full-sentence MSD = –2.47), which would predict no perceptual shift whatsoever.

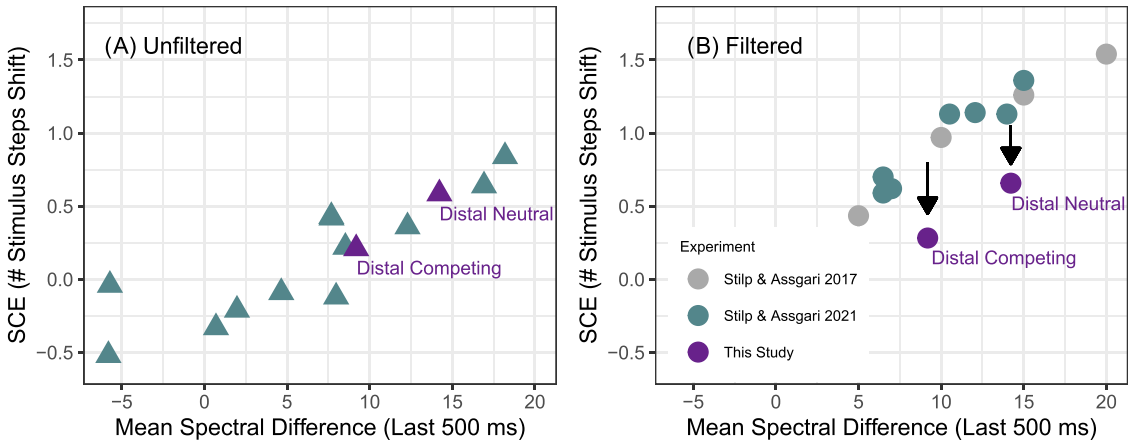

FIG. 3. (Color online) SCE magnitudes as functions of the MSDs for the last 500 ms (proximal window) of context sentences in the present study (darkest/purple symbols) relative to previous studies of SCEs (lighter/green and grey symbols; see legend). The arrows in panel (B) indicate that the SCEs produced by filtered sentences of the present study are markedly smaller than those produced by other sentences with similar MSDs in previous studies.

Instead, categorization shifts were larger and in line with (and in the contrastive direction of) spectral contents of the Proximal windows. Across Stilp and Assgari (2021) and the present study, both temporally nonadjacent (Distal MSDs) and cumulative spectral properties (full-sentence MSDs) made different (and ultimately inaccurate) predictions than temporally adjacent (Proximal MSDs) context as to how perception would be influenced by preceding context.

The present study was designed as an explicit test of the finding from Stilp and Assgari (2021) that the proximal portion of context sentences biased target consonant categorization much more than distal portion did. Therefore, direct comparison of results across studies illustrates the impact of deliberately versus incidentally presenting unfiltered sentence stimuli whose proximal windows have certain spectral emphases. For this comparison, MSDs were recalculated over the last 500 ms of the Stilp and Assgari context sentences to match the durations of Proximal windows in the present stimuli. SCEs produced by unfiltered context sentences in the present study align extremely well with those following other unfiltered sentences in Stilp and Assgari (2021) [Fig. 3(A)]. The present results in filtered conditions were also compared with SCEs from other studies that used the same sentence token (Stilp and Assgari, 2017, 2021). In previous studies, a particular spectral region of the context sentence was amplified uniformly throughout its entire duration; thus, those stimuli lacked competing or even neutral spectral information. Here, the filtered sentences possessed abrupt changes in MSD properties across distal and proximal windows in order to pattern after MSD properties in unfiltered sentences. These abrupt changes in filtered sentence MSDs resulted in markedly smaller SCEs than those reported in previous studies [Fig. 3(B)]. Figure 3 also sheds further light on SCE magnitudes being larger in Distal Neutral conditions than Distal Competing conditions. One explanation for this result is that contradictory spectral information in the Distal windows weakened the spectral bias of the Proximal windows and how they influenced target sound categorization. Alternatively, the stimuli in the Distal Neutral conditions concurrently possessed stronger MSDs in their Proximal windows. Given the overarching relationship that context sentences with stronger MSDs produce larger SCE magnitudes (Stilp and Assgari, 2017, 2021) as illustrated in Fig. 3, it is difficult to distinguish which factor is primarily responsible for this pattern of results.

One surprising result was the lack of a significant difference in SCE magnitudes across filtered and unfiltered blocks (the nonsignificant Spectral Peak by Condition interaction). This diverges from clear patterns of filtered context sentences producing larger SCEs than unfiltered sentences [Stilp and Assgari, 2019, 2021; compare green shapes in Figs. 3(A) and 3(B)]. These previous studies observed this trend across many blocks and experiments, whereas the present study measured four SCEs (two following unfiltered context sentences, and two following filtered sentences) in one experiment from one group of participants. However, differences in stimulus construction across studies merit consideration. Previous studies used filters to modify the MSD of *entire* context sentences, then compared their resulting SCEs to those produced by unfiltered context sentences where the *proximal* windows were most influential on subsequent categorization behavior. Here, filtered and unfiltered sentence stimuli were more alike in that MSDs of the proximal windows were matched. Future research will elucidate whether the null result observed here was a spurious finding or that it provides insight as to how more comparable construction of filtered and unfiltered sentence stimuli yields comparable influences on subsequent speech categorization.

Speech acoustics are notoriously variable from moment to moment and from sound to sound. Previous efforts to study how reliable spectral properties of a listening context inform subsequent perception made these acoustics relatively uniform over time (e.g., using filters to amplify a prescribed frequency region, presenting a sequence of pure tones within a prescribed frequency region). Such approaches resulted in distal and proximal windows of the context whose spectral and/or temporal properties were in agreement (e.g., a spectral peak in low-F_3 frequencies

throughout the entire context). Such persistent regularities may be broadly prevalent in listening environments (e.g., filtering imposed by a particular medium such as a loudspeaker, reverberation characteristics of a given room), but similar regularities are not prevalent to the same degree in speech. While some regularities might emerge on longer time scales (such as those imposed by a given talker's vocal tract length), they become more volatile on shorter time scales as a function of which sounds are being produced at that moment in time. The present approach capitalized on this by presenting stimuli that did or did not have relatively prominent spectral peaks across slightly longer time scales (competing or neutral spectral properties in the distal window, respectively) before changing abruptly due to the sounds being produced at the end of the sentence (in the proximal window). These context sentences influenced the categorization of the subsequent speech target, primarily due to the spectral characteristics of the proximal window. This highlights how recent acoustic context (and recent perceptual experience in general) is an evolving basis of comparison. Perceivers' future behavior is efficiently informed by operating on multiple time scales in parallel. Different sounds will have distal and proximal windows whose spectrotemporal properties agree or disagree, so perceivers must remain flexible in how they utilize and weight recent experience.

ACKNOWLEDGMENTS

The authors wish to thank Isabel Adames, Ella Beilman, Kate Criner, Betsy Sellers, Pratistha Thapa, Sydney Tharp, Maaike Van der Veer, and Sara Wardip for assistance with data collection. This work was supported by National Institutes of Health, National Institute on Deafness and Other Communication Disorders, Grant. No. R01 DC020303.

Ainsworth, W. A. (**1975**). "Intrinsic and extrinsic factors in vowel judgments," in *Auditory Analysis and Perception of Speech* (Elsevier, Amsterdam), pp. 103–113.

Bates, D. M., Maechler, M., Bolker, B., and Walker, S. (**2014**). "lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7," available at https://cran.r-project.org/web/packages/lme4/index.html (Last viewed April 11, 2023).

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1990**). "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [CD-ROM]" (NIST, Gaithersburg, MD).

Green, P., and MacLeod, C. J. (**2016**). "SIMR: An R package for power analysis of generalized linear mixed models by simulation," Methods Ecol. Evol. **7**(4), 493–498.

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

Heffner, C. C., Newman, R. S., and Idsardi, W. J. (**2017**). "Support for context effects on segmentation and segments depends on the context," Atten. Percept. Psychophys. **79**(3), 964–988.

Holt, L. L. (**2006**). "The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization," J. Acoust. Soc. Am. **120**(5), 2801–2817.

Kidd, G. R. (**1989**). "Articulatory-rate context effects in phoneme identification," J. Exp. Psychol. Hum. Percept. Perform. **15**(4), 736–748.

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

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

The MathWorks Inc. (**2021**). *MATLAB (No. R2021a)* (The Mathworks, Inc., Natick, MA).

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**(5), 2088–2113.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099.

R Core Team (**2021**). "R: A language and environment for statistical computing," https://www.r-project.org/ (Last viewed April 11, 2023).

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

Reinisch, E., Jesse, A., and McQueen, J. M. (**2011**). "Speaking rate from proximal and distal contexts is used during word segmentation," J. Exp. Psychol. Human Percept. Perform. **37**(3), 978–996.

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., and Cook, S. (**2012**). "Development and validation of the AzBio sentence lists," Ear Hear. **33**(1), 112–117.

Stephens, J. D. W., and Holt, L. L. (**2011**). "A standard set of American-English voiced stop-consonant stimuli from morphed natural speech," Speech Commun. **53**(6), 877–888.

Stilp, C. E. (**2020**). "Acoustic context effects in speech perception," Wiley Interdiscip. Rev. Cogn. Sci. **11**(1), 1–18.

Stilp, C. E., and Assgari, A. A. (**2017**). "Consonant categorization exhibits a graded influence of surrounding spectral context," J. Acoust. Soc. Am. **141**(2), EL153–EL158.

Stilp, C. E., and Assgari, A. A. (**2019**). "Natural speech statistics shift phoneme categorization," Atten. Percept. Psychophys. **81**(6), 2037–2052.

Stilp, C. E., and Assgari, A. A. (**2021**). "Contributions of natural signal statistics to spectral context effects in consonant categorization," Atten. Percept. Psychophys. **83**(6), 2694–2708.

Stilp, C. E., Shorey, A. E., and King, C. J. (**2022**). "Nonspeech sounds are not all equally good at being nonspeech," J. Acoust. Soc. Am. **152**(3), 1842–1849.

Summerfield, Q. (**1981**). "Articulatory rate and perceptual constancy in phonetic perception," J. Exp. Psychol. Hum. Percept. Perform. **7**(5), 1074–1095.

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

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

Watkins, A. J. (**1991**). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortiona," J. Acoust. Soc. Am. **90**(6), 2942–2955.

Winn, M. B., and Stilp, C. E. (**2019**). "Phonetics and the auditory system," in *The Routledge Handbook of Phonetics* (Routledge, Philadelphia, PA), pp. 164–192.