JASA ARTICLE

Parameterizing spectral contrast effects in vowel categorization using noise contexts[a)]

Christian E. Stilp[b)]
Department of Psychological and Brain Sciences, 317 Life Sciences Building, University of Louisville, Louisville, Kentucky 40292, USA

ABSTRACT:
When spectra differ between earlier (context) and later (target) sounds, listeners perceive larger spectral changes than are physically present. When context sounds (e.g., a sentence) possess relatively higher frequencies, the target sound (e.g., a vowel sound) is perceived as possessing relatively lower frequencies, and vice versa. These spectral contrast effects (SCEs) are pervasive in auditory perception, but studies traditionally employed contexts with high spectrotemporal variability that made it difficult to understand exactly when context spectral properties biased perception. Here, contexts were speech-shaped noise divided into four consecutive 500-ms epochs. Contexts were filtered to amplify low-F_1 (100–400 Hz) or high-F_1 (550–850 Hz) frequencies to encourage target perception of /ɛ/ ("bet") or /ɪ/ ("bit"), respectively, via SCEs. Spectral peaks in the context ranged from its initial epoch(s) to its entire duration (onset paradigm), ranged from its final epoch(s) to its entire duration (offset paradigm), or were present for only one epoch (single paradigm). SCE magnitudes increased as spectral-peak durations increased and/or occurred later in the context (closer to the target). Contrary to predictions, brief early spectral peaks still biased subsequent target categorization. Results are compared to related experiments using speech contexts, and physiological and/or psychoacoustic idiosyncrasies of the noise contexts are considered. © *2021 Acoustical Society of America.*
https://doi.org/10.1121/10.0006657

(Received 18 February 2021; revised 3 September 2021; accepted 18 September 2021; published online 15 October 2021)

[Editor: Joshua G Bernstein] Pages: 2806–2816

I. INTRODUCTION

Perception of a given speech sound can be influenced by acoustic properties of surrounding sounds; this fact has been widely documented [for review, see Stilp (2020a)]. Such context effects are not specific to perception of speech, but widespread throughout audition and all perceptual modalities (von Békésy, 1967; Warren, 1985). This pervasiveness is consistent with all perception transpiring in the context of surrounding objects and events in the sensory environment.

The context effect under study in this report is the spectral contrast effect (SCE), where differences in spectral properties between successive sounds (a context sound and a to-be-identified target sound) are perceived to be larger than they actually are. A classic example of this effect was reported by Ladefoged and Broadbent (1957), where a context sentence with emphasized lower-first-formant-frequencies (F_1) increased the number of high-F_1 responses (/ɛ/, "eh" as in "bet") to the target vowel. When the context sentence had higher-F_1 frequencies emphasized, listeners identified the target vowel as the low-F_1 response option (/ɪ/, "ih" as in "bit") more often. In both cases, perception of formant frequencies in the target vowel was shifted away from prominent spectral properties in the preceding context sentence.

In the three-score-plus years since the seminal report of Ladefoged and Broadbent's (1957), much has been learned about the nature of SCEs. They occur on various timescales, from the preceding context stimulus being a sentence [as in Ladefoged and Broadbent (1957)] or just a syllable (Lotto and Kluender, 1998). These effects of timescale have been replicated using nonspeech context stimuli such as sentence-length noise or pure tone sequences (Watkins, 1991; Holt, 2005) or phoneme-length pure tones (Lotto and Kluender, 1998). SCEs are not specific to speech perception, as spectral properties of a brief excerpt of a string quintet can alter categorization of the subsequent musical instrument sound (Stilp *et al.*, 2010). The magnitudes of SCEs in speech and nonspeech perception scale to reflect the degree to which context and target spectra differ (Stilp *et al.*, 2015; Stilp and Assgari, 2017; Frazier *et al.*, 2019). Finally, while SCEs are thought to be low-level in nature (Delgutte, 1996; Delgutte *et al.*, 1996; Stilp, 2020b) and are larger when context and target stimuli are presented ipsilaterally than contralaterally (Watkins, 1991; Holt and Lotto, 2002; Feng and Oxenham, 2018; Stilp, 2020b), their magnitudes can be modulated somewhat by selective attention in competing-talker paradigms (Feng and Oxenham, 2018; Bosker *et al.*, 2020).

In many of these studies, the researchers amplified frequencies in the context sentence that were contrastive with frequencies in the target sound, producing the SCE.

[a)]These results were presented at the 179th Meeting of the Acoustical Society of America (Acoustics Virtually Everywhere).
[b)]Electronic mail: christian.stilp@louisville.edu, ORCID: 0000-0002-5119-201X.

 0001-4966/2021/150(4)/2806/11/$30.00

However, spectral properties of speech change quite rapidly, particularly on the one-second-plus durations typical of sentences. When the researchers added a peak to the context's long-term spectrum, they ceded control as to exactly when that peak occurred in the context and for how long. This obscures key characteristics of when and how earlier sounds influence perception of later sounds. For example, a spectral peak occurring at the end of the context sentence could influence categorization of the subsequent speech target, but its influence would conceivably lessen if the peak occurred earlier in the context sentence (further away in time from the target sound). This cannot be explained by the long-term average spectrum of the context sentence, as that would be essentially unchanged. These notions are supported by reports that SCE magnitudes decreased with increasing silent interstimulus (ISI) intervals separating context and target stimuli (Broadbent and Ladefoged, 1960; Holt and Lotto, 2002; Stilp and Winn, 2021). Studies that held ISI constant and manipulated the timing of spectral peaks in the context produced mixed results. In Holt (2006), when contexts were 2100 ms of pure tones divided into 700-ms epochs in different frequency regions (spectral peak in lower, medium, or higher frequencies) tested in different orders, no systematic context effects emerged for categorization of /da/-/ga/ targets. In Stilp (2018), when contexts were sentences, spectral properties of the last 500 ms of the sentence were more influential than (competing or neutral) spectral properties that preceded those last 500 ms for categorization of /da/-/ga/ targets. These competing findings come from studies that used very different context stimuli, which reinforces the need to systematically manipulate the timing of context spectral properties.

In the present study, key acoustic properties of context sounds were systematically varied to examine their influence on subsequent vowel categorization via SCEs. Speech-shaped noise was used as the context stimulus to control spectrotemporal variability throughout its duration. Three different experimental paradigms were employed to parametrically investigate how SCEs shape speech categorization. In the offset paradigm, spectral peaks began at some point during the context but terminated at its end. In the onset paradigm, spectral peaks began at context onset and continued for some duration. Finally, in the single paradigm, fixed-duration spectral peaks occurred at different temporal positions in the context. The gain of filters that added spectral peaks to the noise context was also varied in each paradigm to elucidate interactions between spectral peak magnitude and timing.

Neural adaptation has been proposed to be the primary mechanism underlying SCEs [for discussions, see Delgutte (1996), Delgutte *et al.* (1996), and Stilp (2020b)]. By this mechanism, frequencies in the context adapt neurons coding those frequencies, making them less responsive when the target sound is introduced. Neurons coding neighboring frequencies would be unadapted/less adapted and thus relatively more responsive to the frequencies in the target, producing a neural contrast consistent with perception of shifted spectral content in the target. While neural adaptation occurs throughout the auditory system, Stilp (2020b) suggested that peripheral adaptation is primarily (but not exclusively) responsible for producing SCEs. In those studies, stimulus presentation that favored peripheral processing (context and target presented monaurally) produced the largest SCEs whereas stimulus presentation that favored central processing (context presented to one ear followed by target presented to the contralateral ear) produced small but still significant SCEs. This suggestion is also consistent with reports of SCE magnitudes decreasing as the duration of the silent inter-stimulus interval (ISI) between context and target stimuli increased (Broadbent and Ladefoged, 1960; Holt and Lotto, 2002; Sjerps *et al.*, 2018; Stilp and Winn, 2021).

To the extent that the correct mechanism and processing (i.e., peripheral versus central contributions) underlying SCEs have been identified, corroborating evidence should be obtainable without large changes to stimulus (ear of) presentation or duration (either of the context stimulus itself or of stimulus ISIs). Here, stimulus duration, interstimulus interval, and ears of presentation were held constant while manipulating the duration and timing of spectral peaks in the context. Two main hypotheses motivated the present investigation. First, longer-duration spectral peaks in the context are hypothesized to produce larger SCEs. The time constants of neural adaptation generally increase along the ascending auditory pathway [but see Pressnitzer *et al.* (2008)], so longer-duration peaks that elicit adaptation at progressively more central levels of this pathway are predicted to result in larger SCE magnitudes. This hypothesis draws from the results of Holt (2006), where longer sequences of pure-tone contexts produced larger SCEs in /d/-/g/ categorization. Second, SCE magnitudes are hypothesized to diminish as the temporal interval between the context spectral peak and the target increases. Contributions from peripheral neural adaptation would decrease with this increasing temporal nonadjacency, as demonstrated for increases in the ISI between context and target stimuli (Broadbent and Ladefoged, 1960; Holt and Lotto, 2002; Stilp and Winn, 2021).

Specific predictions varied by experimental paradigm. In the offset paradigm, all blocks were predicted to produce SCEs, with SCE magnitudes increasing as the duration of the context spectral peak increases. Longer spectral peak durations were predicted to produce larger SCEs in the onset paradigm as well. But, given that the spectral peak was often separated from the target vowel by spectrally neutral context (lacking any added spectral peaks to bias responses), SCEs were predicted to be extinguished in at least the most extreme case (where the spectral peak occupies the first 500 ms of the context followed by 1500 ms of spectrally neutral speech-shaped noise). In the single paradigm, SCEs were predicted to be extinguished in most blocks owing to the fixed duration of the spectral peak and the temporal separation of the spectral peak from the context offset by spectrally neutral noise. Across paradigms, offset SCEs were predicted to be larger than onset SCEs, as consistently

presenting the spectral peak just before the target (in the offset paradigm) was expected to produce larger SCEs than when the spectral peak was separated from the target by spectrally neutral context (in the onset paradigm). Single SCEs were predicted to be smallest overall, as restricting the spectral peak duration to 500 ms was expected to produce smaller SCEs than when peak durations were longer (up to 2000 ms in the offset and onset paradigms).

II. METHODS

A. Participants

Thirty-one listeners participated in the experiment. All were recruited using flyers on the University of Louisville campus. All reported that English was their native language and that they had no known hearing impairments. Listeners were compensated for their participation at a rate of $10/h. Of these 31 listeners, 17 met all eligibility criteria and completed the experiment.

B. Stimuli

1. Contexts

Filtered-noise contexts were generated according to the following procedure. First, for each context stimulus, 2 two-second samples of speech-shaped noise were created by filtering white (random) noise to produce a spectrum that was flat up to 500 Hz then decreased at a rate of –9 dB/octave beyond that point. Both samples were set to a constant root mean square (RMS) amplitude. Next, one of the noise tokens was processed by a 300-Hz-wide filter in one of two frequency regions near F_1 in the target vowels. The low-F_1 region spanned 100–400 Hz (just below F_1 in the /ɪ/ endpoint) and the high-F_1 region spanned 550–850 Hz (just above F_1 in the /ɛ/ endpoint); each filter had 50-Hz transition regions. Given the low center frequency of the low-F_1 filter, the slope on its low-frequency side (below 100 Hz, ≈ –8 dB/octave) was much shallower than the slope on its high-frequency side (above 400 Hz, ≈ –66 dB/octave); slopes were more comparable for the high-F_1 filter (below 550 Hz, ≈ –80 dB/octave; above 850 Hz, ≈ –135 dB/octave). The level of filter gain in the passband (with 0 dB gain at other frequencies) was either +10 or +20 dB. This created "low-F_1-amplified" and "high-F_1-amplified" versions of the context. All filters were created using the fir2 function in MATLAB (MathWorks, Inc., Natick, MA) with 1200 coefficients. Portions of these two noise samples were excised and concatenated to create a context stimulus with a finite-duration spectral peak occurring at a specific temporal position. Novel noise tokens were generated each time so that no participant heard any context stimulus more than once.

Three experimental paradigms were tested. In the offset paradigm, the spectral peak started at some point during the context but continued until its end. Context stimuli in offset trials possessed a spectral peak for the final 500, 1000, 1500, or entire 2000 ms [Fig. 1(A)]. In the onset paradigm, the spectral peak started at the onset of the context but had variable durations. Context stimuli in onset trials possessed a spectral peak for the initial 500, 1000, 1500, or entire 2000 ms [Fig. 1(B)]. In the single paradigm, a fixed-duration 500-ms spectral peak occurred at different temporal positions within the context. Context stimuli in single trials possessed a spectral peak from 0–500, 500–1000, 1000–1500, or 1500–2000 ms [Fig. 1(C)]. Some conditions were repeated across paradigms (cf. Fig. 1: the bottom figures for offset and onset; the top figures for onset and single; the top figure of offset and the bottom figure of single). This was done to facilitate comparisons to other (related) conditions in a given paradigm by testing them all in the same session. The beginning and end of the context, as well as all transitions between unfiltered and filtered noise segments, were processed by 2-ms cosine-squared ramps. Finally, speech-shaped noise contexts without added spectral peaks were also generated for use in a baseline categorization task. All contexts were low-pass filtered with a cutoff frequency of 5000 Hz, matching the spectral bandwidth of the target vowels.

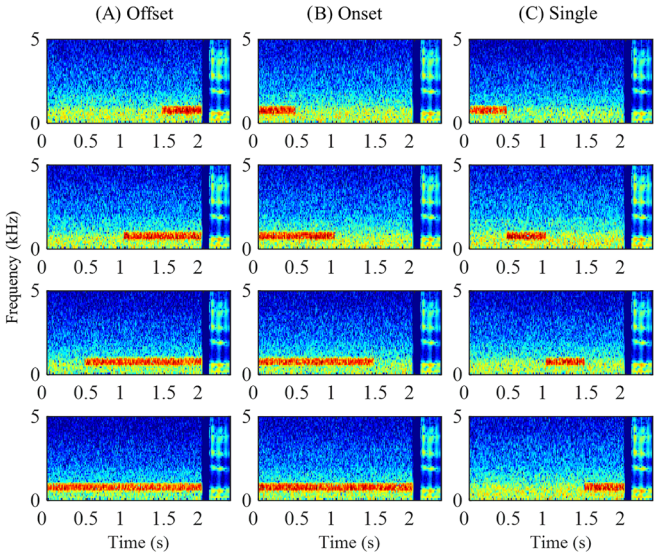

FIG. 1. (Color online) Sample trials in each experimental paradigm. The two-second context was divided into four successive 500-ms epochs, during which frequencies that produced the SCE were amplified. The sample trials illustrated here show +20 dB amplification of high-F_1 frequencies in the context (550–850 Hz). The 246-ms target vowel was presented following a silent 50-ms interstimulus interval.

2. Targets

Target vowels were the same /ɪ/-to-/ɛ/ continuum as previously tested by Stilp and colleagues [e.g., Stilp *et al.* (2015)]. For a detailed description of the generation procedures, see Winn and Litovsky (2015). Briefly, tokens of /ɪ/ and /ɛ/ were spoken and recorded by the author. Formant contours were extracted from each token and slightly modified in PRAAT (Boersma and Weenink, 2014). In the /ɪ/ endpoint, F_1 linearly increased from 400 to 430 Hz while F_2 linearly decreased from 2000 to 1800 Hz. In the /ɛ/

https://doi.org/10.1121/10.0006657

endpoint, F_1 linearly decreased from 580 to 550 Hz while F_2 linearly decreased from 1800 to 1700 Hz. These F_1 trajectories were linearly interpolated to create a ten-step continuum of formant tracks; linear interpolations were also performed for F_2 trajectories. A single voice source was extracted from the /ɪ/ endpoint. Formant tracks were used to filter this source, producing the ten-step continuum of vowel tokens. For all vowels, energy above 2500 Hz was replaced with the energy high-pass-filtered from the original /ɪ/ token. Stimuli were 246 ms in duration with fundamental frequency set to 100 Hz throughout the vowel. Vowels were set to the same RMS amplitude as the unfiltered noise.

Experimental trials were created by concatenating one noise context and one vowel target with a 50-ms silent interstimulus interval. Final presentation levels of the noise context varied depending on spectral-peak duration and filter gain. For trials featuring +20 dB spectral peaks, average presentation levels ranged from 73 (500-ms spectral-peak duration) to 81 dB SPL (spectral peak during the entire context). For trials featuring +10 dB spectral peaks, average presentation levels ranged from 66 (500-ms spectral-peak duration) to 71 dB SPL (spectral peak during the entire context).[1] Average presentation levels for the target vowels and the noise context without the added spectral peak was 64 dB SPL.

C. Procedure

During their initial visit to the lab, 31 listeners provided informed consent and participated in three screening tasks. First, listeners completed an abridged version of the LEAP-Q (Marian *et al.*, 2007) to assure that their native language was English. Eligibility was determined by responding that English was one's dominant language (question 1), acquired first (question 2), the language to which participants were exposed to more than any other (question 3), and used as the primary read (question 4) and spoken language (question 5). Three participants did not meet this eligibility criterion and did not proceed in the experiment. Second, listeners sat at a personal computer inside a sound-attenuating booth (Acoustic Systems, Inc., Austin, TX) to complete 48 practice trials: each endpoint of the vowel target series (/ɪ/, /ɛ/) was presented in one trial in each of the twelve trial types illustrated in Fig. 1 at each level of filtering (low-F_1 frequencies or high-F_1 frequencies amplified by +20 dB). After each trial, participants clicked the mouse to indicate whether the target vowel sounded more like "ih (as in 'bit')" or "eh (as in 'bet')." 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) before being presented over circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY). A custom MATLAB script led the participants through the session, and feedback was provided on each trial. Eligibility was determined by categorizing the vowels with at least 80% accuracy. If the listener did not meet this criterion, s/he repeated the practice session up to two more times as needed. All listeners met this criterion. Third, listeners participated in a hearing screening to assure that pure-tone thresholds in each ear were ≤20 dB hearing level (HL) at octave frequencies from 125 to 8000 Hz (ANSI, 2010). The experimenter conducted the screening on a Maico MA27 audiometer. Six participants did not meet the criterion for the hearing screening; they did not proceed in the experiment. In all, the initial visit took approximately 35 min.

Listeners who passed all screenings proceeded to the main experiment, which was fully within-subjects and consisted of six one-hour sessions. Comparing SCE magnitudes across different conditions and across different paradigms was a key goal for this study, which made a long within-subjects experiment preferable to shorter between-subjects experiments. At the beginning of the first session, listeners were oriented to the task by completing 80 trials where each target vowel was presented eight times following a speech-shaped noise context without any added spectral peaks. Each session then tested one of the three experimental paradigms (offset, onset, single) at one of the two levels of filter gain (+10 dB, +20 dB). Given the prohibitive nature of testing all possible orders of these sessions (six factorial = 720 possible orders), they were tested in pseudo-random orders such that after each six participants, each condition had been tested in each position once. Sessions were all scheduled for different days to avoid fatigue. Each session consisted of 640 trials: eight repetitions of each target vowel (10) and each filtering condition (2; low-F_1-amplified and high-F_1-amplified) in each temporal arrangement (4; rows in Fig. 1) at one level of filter gain. These trials were divided into blocks of 160 trials (two repetitions of every trial type), between which listeners took self-paced breaks. In all, each session lasted approximately 50 min, with the first session lasting approximately 55 min.

III. RESULTS

Of the 22 listeners who passed all screeners, five elected to withdraw participation before completing the experiment, resulting in 17 listeners who provided complete datasets for statistical analyses. Data were analyzed in mixed-effects models using the lme4 package (Bates *et al.*, 2014) in R (R Development Core Team, 2021).

A. Within-paradigm analyses

Responses were divided by paradigm and analyzed in separate mixed-effects models. Responses were transformed using the binomial logit linking function. The dependent variable was modeled as binary ("ih" or "eh" responses coded as 0 and 1, respectively). Fixed effects in the model included: target (coded as a continuous variable from 1 to 10 then mean-centered), filter (sum coded; high $F_1 = -0.5$, low $F_1 = +0.5$), gain (coded as a continuous variable then mean-centered), epoch (dummy coded with epoch 4 as the reference level[2]), and all possible interactions. Random

slopes were included for each main fixed effect that significantly improved model fit, and a random intercept of listener was also included in each model. All models were run using the bobyqa optimization with a maximum of 800 000 iterations.

In addition to the omnibus models, SCEs were calculated for each listener in each epoch of each condition following established procedures (Stilp *et al.*, 2015). Each listener's responses in a given block were fit with a logistic regression with fixed effects of target, filter, and their interaction. The 50% points were identified on the logistic regression fits to responses following low-F_1-amplified contexts and high-F_1-amplified contexts. These 50% points were then converted into the stimulus step number that listeners would label as "eh" 50% of the time. Vowel targets were numbered from 1 to 10, so this stimulus number was interpolated as needed. The SCE was measured as the number of stimulus steps separating these 50% points. These are illustrated to accompany within-paradigm analyses and utilized directly in forthcoming across-paradigm analyses.

Figure 2 depicts SCEs in the offset paradigm as functions of epoch and filter gain. The mixed-effects model fitted to these responses had fixed main effects listed above (target, filter, gain, epoch, and all possible interactions), as did the models analyzing results in the onset and single paradigms, random slopes for target and gain, and random intercepts for listeners (supplementary Table I[3]). As expected, the main effect of target was significant ($Z = 17.694$, $p < 2e-16$), indicating that listeners were more likely to respond "eh" with each rightward step along the vowel target continuum

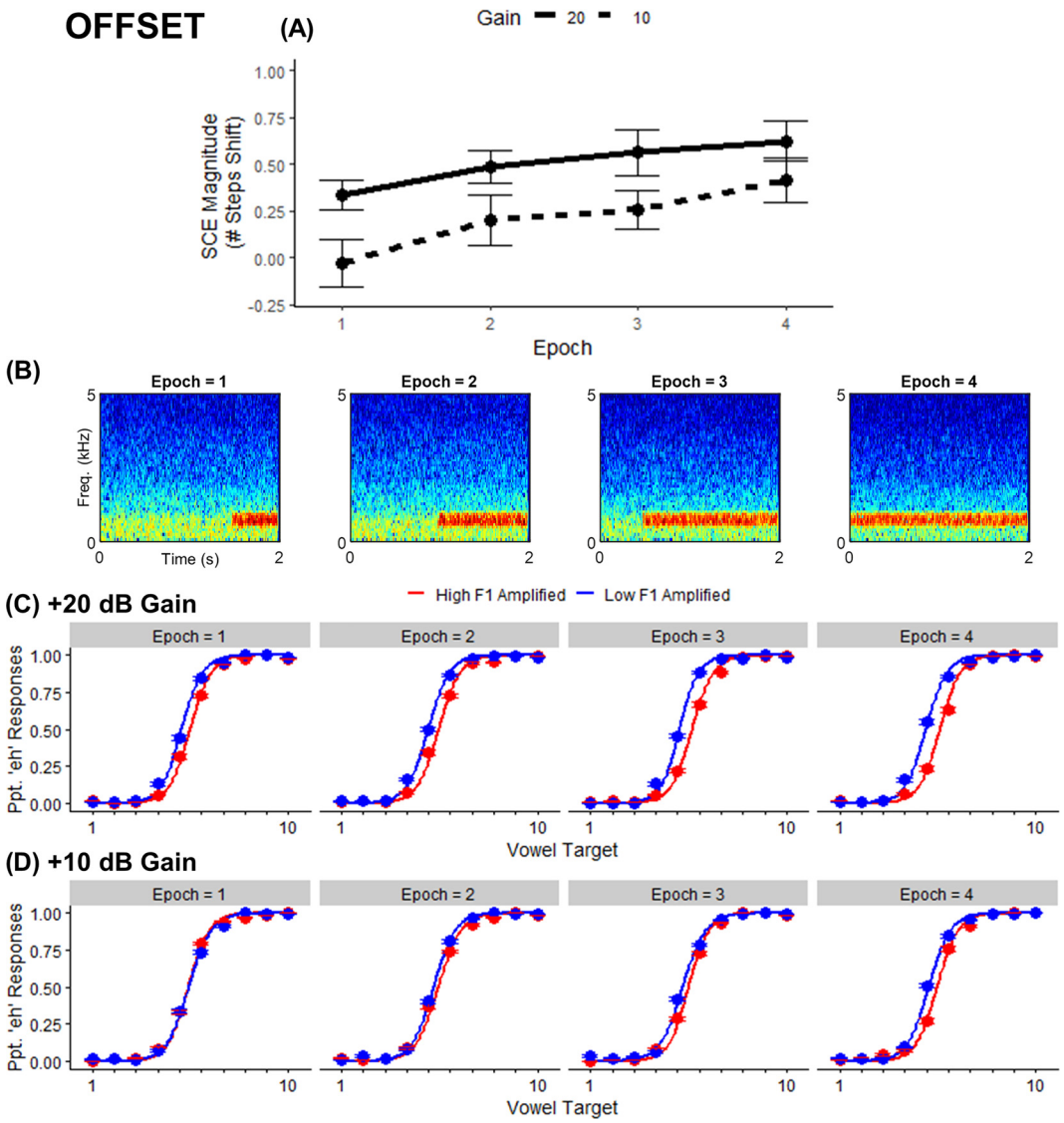

FIG. 2. (Color online) Results from the offset paradigm. (A) Mean SCE magnitudes are plotted as functions of context epoch (abscissa) and filter gains (lines). SCEs were measured as the number of stimulus steps separating 50% points on response functions following low-F_1-amplified and high-F_1-amplified contexts. Error bars depict ±1 standard error of the mean. (B) Spectrograms illustrating the context stimuli presented in each epoch (example contexts with high-F_1 frequencies amplified by +20 dB shown). (C) Psychometric functions in each epoch generated by the mixed-effects model fit to "eh" responses following contexts with +20 dB spectral peaks. Blue lines depict responses to vowels following low-F_1-amplified contexts; red lines depict responses to vowels following high-F_1-amplified contexts. Circles depict mean proportions of "eh" responses; error bars depict ±1 standard error of the mean. (D) The same as (C) but for +10 dB gain conditions.

https://doi.org/10.1121/10.0006657

(toward the /ɛ/ endpoint). Also, as expected, the main effect of filter was significant ($Z = 7.714$, $p < 2e-14$), as listeners responded "eh" more often when the noise context possessed a spectral peak in the low-F_1 frequency region compared to the high-F_1 frequency region [cf. horizontal separation between functions in Figs. 2(C) and 2(D)]. This is consistent with the predicted direction of SCEs. The interaction between filter and gain was positive and significant ($Z = 2.069$, $p = 0.039$), indicating that SCE magnitudes were larger following +20 dB amplification of spectral peaks in the context compared to +10 dB amplification [larger separation between functions in Fig. 2(C) than Fig. 2D; also the solid line in Fig. 2(A) being higher on average than the dashed line]. Most importantly, in the filter-by-epoch-number interactions (which average across the two levels of filter gain), with epoch 4 serving as the reference level, SCEs were not diminished in epoch 3 ($Z = -0.827$, $p = 0.408$) but were significantly smaller in epoch 2 ($Z = -2.014$, $p = 0.044$) and epoch 1 ($Z = -3.859$, $p = 0.001$). While the filter-by-gain interaction and two filter-by-epoch-number interactions were statistically significant, none of the three-way interactions between filter, gain, and epoch number were significant (all $Z < 0.877$, $p > 0.381$), indicating that the effect of gain on SCE magnitudes observed in epoch 4 was similar compared to other epochs. Additional results from the model included a slight bias to respond "eh" more often overall (52.00%; significant intercept), and fewer "eh" responses in epoch 3 compared to epoch 4 (main effect of epoch 3; see supplementary Table I[3] for details).

Figure 3 depicts SCEs in the onset paradigm as functions of epoch and filter gain. The mixed-effects model fit to

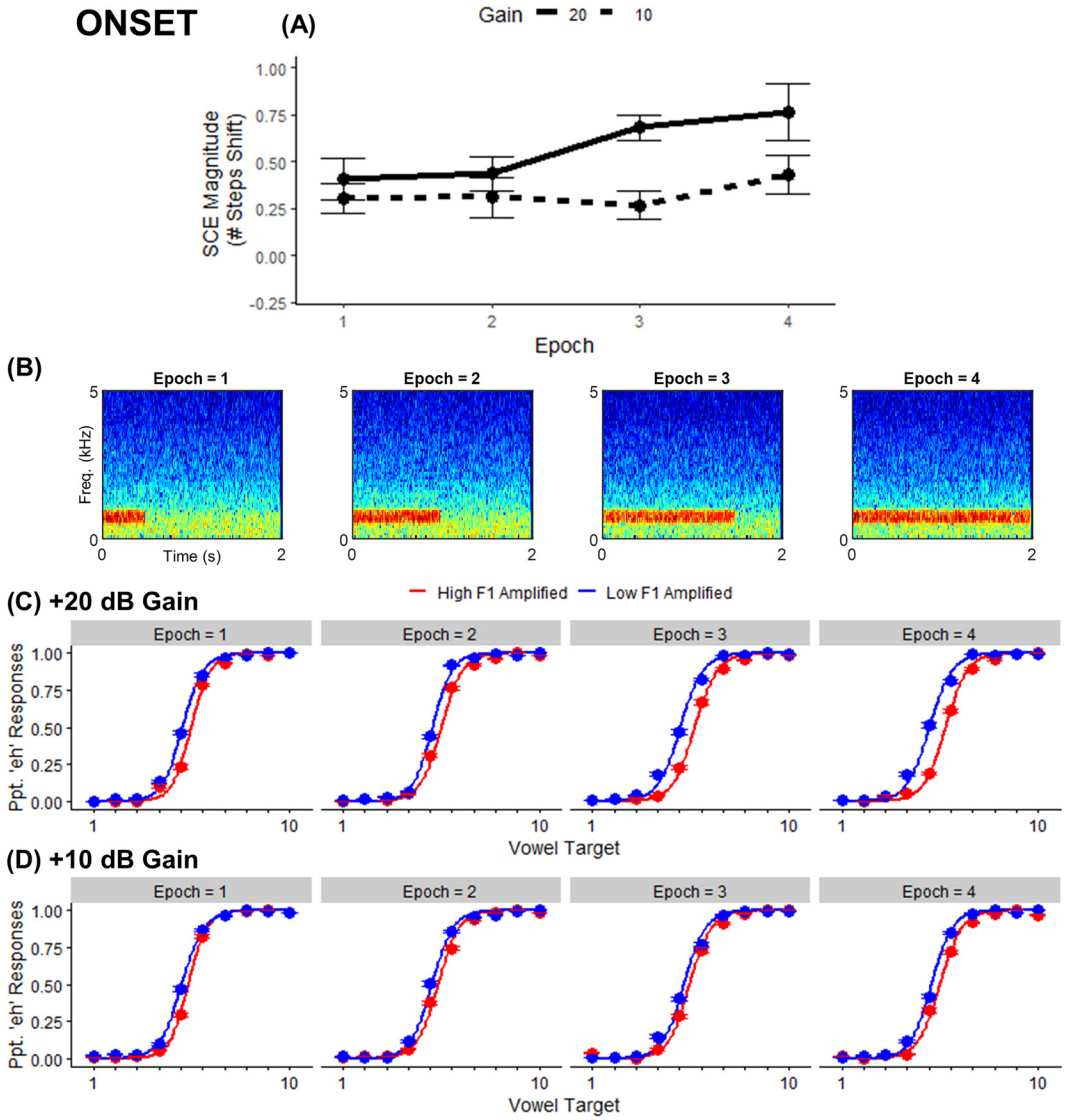

FIG. 3. (Color online) Results from the onset paradigm. (A) Mean SCE magnitudes are plotted as functions of context epoch (abscissa) and filter gains (lines). SCEs were measured as the number of stimulus steps separating 50% points on response functions following low-F_1-amplified and high-F_1-amplified contexts. Error bars depict ±1 standard error of the mean. (B) Spectrograms illustrating the context stimuli presented in each epoch (example contexts with high-F_1 frequencies amplified by +20 dB shown). (C) Psychometric functions in each epoch generated by the mixed-effects model fit to "eh" responses following contexts with +20 dB spectral peaks. Blue lines depict responses to vowels following low-F_1-amplified contexts; red lines depict responses to vowels following high-F_1-amplified contexts. Circles depict mean proportions of "eh" responses; error bars depict ±1 standard error of the mean. (D) The same as (C) but for +10 dB gain conditions.

these responses had the same fixed-effects structure as that detailed for the offset paradigm, with random slopes for target, filter, and gain, as well as random intercepts for listeners (supplementary Table II[3]). As observed with the offset paradigm, the main effects of target ($Z = 13.431$, $p < 2e-16$), filter ($Z = 8.311$, $p < 2e-16$), and the filter-by-gain interaction ($Z = 2.440$, $p = 0.015$) were all significant and in the predicted directions [larger separation between functions in Fig. 3(C) than Fig. 3(D); solid line in Fig. 3(A) being higher on average than dashed line]. Most importantly, results again patterned similarly to the offset data in that SCEs in epoch 4 were not larger than those observed in epoch 3 ($Z = -1.284$, $p = 0.199$), but were larger than those observed in epoch 2 ($Z = -2.195$, $p = 0.028$) and epoch 1 ($Z = -2.609$, $p = 0.009$). Also, like the offset data, no three-way interactions between filter, gain, and epoch number were significant (all $Z < 0.412$, $p > 0.402$). Additional results from the model included more "eh" responses in epoch 2 and epoch 1 relative to epoch 4 (main effects of these epochs), and changes in psychometric function slopes in epoch 1 compared to epoch 4 (target-by-epoch 1 interaction, target-by-filter-by-epoch 1 interaction; see supplementary Table II[3] for details).

Figure 4 depicts SCEs in the single paradigm as functions of epoch and filter gain. The mixed-effects model fitted to these responses had the same fixed-effects structure as other models, with random slopes for target, filter, and gain, as well as random intercepts for listeners (supplementary Table III[3]). The

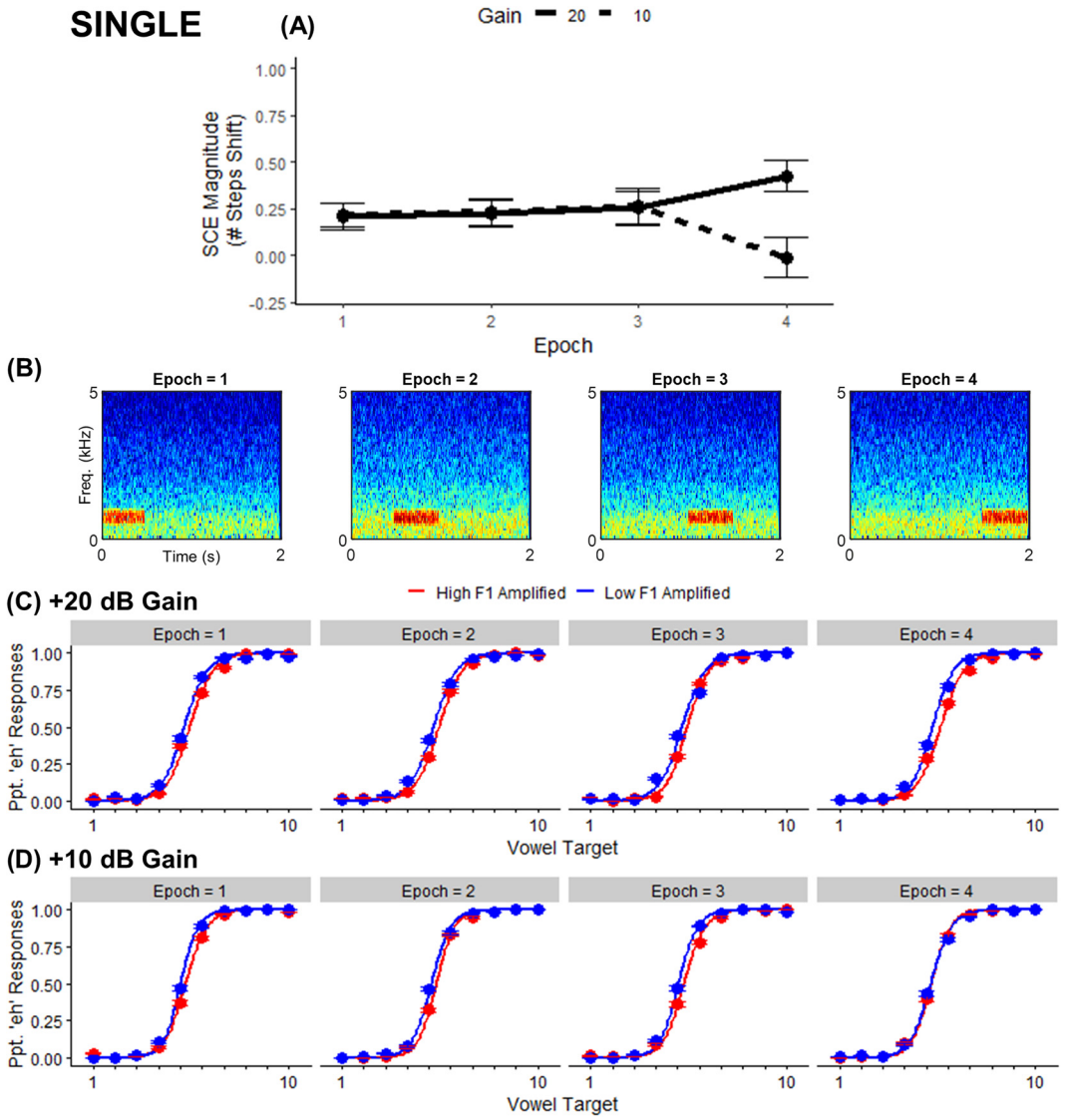

FIG. 4. (Color online) Results from the single paradigm. (A) Mean SCE magnitudes are plotted as functions of context epoch (abscissa) and filter gains (lines). SCEs were measured as the number of stimulus steps separating 50% points on response functions following low-F_1-amplified and high-F_1-amplified contexts. Error bars depict ±1 standard error of the mean. (B) Spectrograms illustrating the context stimuli presented in each epoch (example contexts with high-F_1 frequencies amplified by +20 dB shown). (C) Psychometric functions in each epoch generated by the mixed-effects model fit to "eh" responses following contexts with +20 dB spectral peaks. Blue lines depict responses to vowels following low-F_1-amplified contexts; red lines depict responses to vowels following high-F_1-amplified contexts. Circles depict mean proportions of "eh" responses; error bars depict ±1 standard error of the mean. (D) The same as (C) but for +10 dB gain conditions.

main effects of target ($Z = 18.116$, $p < 2e-16$), filter ($Z = 2.675$, $p = 0.007$), and the filter-by-gain interaction ($Z = 2.828$, $p = 0.005$) were again all significant and in the predicted directions [larger separation between functions in Fig. 4(C) than Fig. 4(D); solid line in Fig. 4(A) being higher on average than dashed line]. Unlike other paradigms, no filter-by-epoch-number interactions were statistically significant (all $Z < 1.266$, $p > 0.205$). The Filter-by-gain-by-epoch-number interactions indicated that filter gain modified SCE magnitudes differently in epoch 4 than in any other epoch (filter-by-gain-by-epoch 3: $Z = -2.673$, $p = 0.008$; filter-by-gain-by-epoch 2: $Z = -2.127$, $p = 0.033$; filter-by-gain-by-epoch 1: $Z = -2.750$, $p = 0.006$). As illustrated in Fig. 4(A), filter gain did not affect SCE magnitudes in epochs 1–3 but did in epoch 4. Additional results from the model included fewer "eh" responses and shallower psychometric function slopes as filter gain increased (negative main effect of gain, negative target-by-gain interaction), more "eh" responses in epoch 1 relative to epoch 4 (main effect of epoch 1), and shallower psychometric function slopes in epochs 3 and 1 for the filter-by-gain interaction compared to epoch 4 (negative target-by-filter-by-gain-by-epoch number interactions; see supplementary Table III[3] for details).

B. Across-paradigm analyses

To facilitate comparisons of SCEs across the three experimental paradigms, SCEs were first calculated for each listener by fitting logistic regressions to their responses in each epoch of each paradigm [as described above and illustrated in Figs. 2(A), 3(A), and 4(A)]. These SCEs served as the dependent measure in linear mixed-effects regressions. To test the global across-paradigm predictions laid out in Sec. I, the ideal fixed-effects structure would factor-code the fixed effect of paradigm while testing all other fixed effects at their means (i.e., continuously coded) rather than at specific levels (i.e., factor coded). This is straightforward for the fixed effect of gain but less so for epoch. For the sake of simplicity, epoch number (1, 2, 3, 4) was mean-centered in the model (n.b., results are consistent with epoch being omitted from the model altogether). All interactions between fixed effects were also included in the model. The random effects structure was determined by starting with the minimal model (random intercepts for listeners) and adding random slopes for fixed effects incrementally. Random slopes for gain significantly improved model fit, but other random slopes failed to converge and thus were not included in analyses. Contrary to predictions, offset SCEs (mean of by-listener SCEs, averaging across epoch and filter gain- $= 0.35$ stimulus steps) were not larger than onset SCEs (mean SCE $= 0.45$ stimulus steps); they were in fact smaller ($t = -1.99$, $p = 0.047$). Consistent with predictions, single SCEs (mean $= 0.23$ stimulus steps) were significantly smaller than offset SCEs ($t = -2.74$, $p = 0.007$) as well as onset SCEs ($t = -4.73$, $p = 4e-6$). Additional analyses of individual epochs across paradigms are provided as supplementary material.[3]

IV. DISCUSSION

Many studies have reported acoustic context effects in speech perception using speech contexts [see Stilp (2020a) for review]. However, the extreme acoustic variability of speech complicates interpretations of the relevant temporal characteristics of the context driving these effects. In the present investigation, spectral peaks were added to a speech-shaped noise context stimulus to control spectrotemporal variability. Two main hypotheses were tested: (1) longer durations for the spectral peak in the context would produce larger SCEs and (2) longer temporal intervals between the context spectral peak and target vowel would diminish SCE magnitudes. Specifically, SCEs in the offset paradigm were predicted to grow as the duration of the context spectral peak increased (according to hypothesis 1). This pattern was also predicted for the onset paradigm (also according to hypothesis 1), but with smaller SCE magnitudes compared to the offset paradigm due to spectrally neutral noise separating the spectral peak from the target (according to hypothesis 2). SCEs in the single paradigm were predicted to be smallest overall owing to the fixed duration of the spectral peak (hypothesis 1) and the temporal separation of the peak from the target (hypothesis 2). Consistent with hypothesis 1, SCE magnitudes increased with increasing spectral peak durations in the offset and onset paradigms, and these SCEs were significantly larger than those produced in the (fixed-peak-duration) single paradigm. Contrary to hypothesis 2, SCEs in the offset paradigm were not larger than those in the onset paradigm, and temporal separations of context spectral peaks and target vowels in onset and single paradigms did not extinguish SCEs.

All experimental paradigms produced two expected results. First, adding spectral peaks to the context at frequencies that were contrastive with frequencies in the target vowels produced SCEs (significant main effects of filter). Second, larger filter gains (+20 dB compared to +10 dB) added larger spectral peaks to the context, which in turn generally produced larger SCEs (significant filter-by-gain interactions). This latter relationship has been observed for speech and nonspeech contexts as well as speech and nonspeech targets (Stilp *et al.*, 2015; Stilp and Assgari, 2017; Frazier *et al.*, 2019). Beyond these shared effects, results from the offset paradigm were the most straightforward (Fig. 2). Consistent with predictions, as the duration of the context spectral peak increased, SCE magnitudes increased (filter-by-epoch-number interactions). This supports the first hypothesis, as shorter-duration spectral peaks would be expected to elicit adaptation in neurons closer to the auditory periphery (where time constants of adaptation are generally shorter) but not neurons in the central auditory system (where adaptation time constants are generally longer). As spectral peak duration increases, peripheral and central auditory neurons would adapt and thus both contribute toward the SCE.[4] The present results refine the conclusions of Holt (2006), who proposed that longer-duration contexts

produced larger SCEs. In that study, contexts were comprised exclusively of sequences of pure tones, thus equating context duration with spectral peak duration. Here, all contexts were set to the same two-second duration, but spectral peak duration varied in the offset and onset paradigms. Thus, longer-duration spectral peaks in the context (which may or may not covary with context duration) produce larger SCEs.

Results in the onset paradigm indicate a more complex relationship between the context spectral peak and subsequent SCE than that observed for the offset paradigm. Longer-duration spectral peaks (epoch 4) produced larger SCEs than shorter-duration peaks (epochs 1 and 2), similar to the offset paradigm (and consistent with hypothesis 1). However, effects of filter gain were only evident in epochs 3 and 4 of the onset paradigm, as opposed to every epoch of the offset paradigm. Additionally, restricting the spectral peak to epoch 1 of the context was predicted to extinguish SCEs, but the effect still occurred, challenging hypothesis 2. This finding might not be surprising considering Broadbent and Ladefoged (1960) reported SCEs occurring despite five or more seconds of silence separating the context sentence from the target vowel [see also Holt and Lotto (2002)]. However, silence and speech-shaped noise are very different acoustic contexts for SCEs (and for auditory perception in general); comparisons across studies when context and target stimuli are separated by silence versus speech-shaped noise should be made with caution. Idiosyncrasies of the noise contexts employed here are reviewed later in this Discussion.

In the single paradigm (Fig. 4), contrary to hypothesis 2, SCEs occurred even when spectrally neutral context separated the spectral peak from the target by 1500 ms. SCE magnitudes were highly comparable across epochs 1–3 (i.e., when spectral peaks were followed by 1500, 1000, or 500 ms of spectrally neutral noise context) and across both levels of filter gain. As in the onset paradigm, neural mechanisms underlying this context effect remained engaged in producing the SCE despite spectrally neutral context separating the spectral peak and the target sound. Results in epoch 4 were the only deviation from this pattern, where SCE magnitudes significantly increased in the +20 dB filter gain condition but were extinguished in the +10 dB gain condition. This latter result is perhaps surprising, considering the spectral peak that was temporally adjacent to the target vowel did not shift listeners' perception but earlier nonadjacent peaks did. This result is not aberrant, as contexts with +10 dB spectral peaks also failed to produce SCEs in epoch 1 of the offset paradigm which is identical to this condition (Fig. 2; see below for possible physiological and/or psychoacoustic explanations why these effects were extinguished).

Results in the present study differed from previous reports of SCEs using speech contexts in several ways. First, adding larger spectral peaks to speech contexts produces systematically larger SCEs (Stilp *et al.*, 2015; Stilp and Assgari, 2017). Here, increasing filter gain produced larger SCEs across all epochs of the offset condition, some epochs of the onset condition, and only the final epoch of the single condition. Second, various speech studies reported that spectral properties at the context offset can produce SCEs (e.g., Lindblom and Studdert-Kennedy, 1967; Mann, 1980; Mann and Repp, 1981; Nearey, 1989; Lotto and Kluender, 1998). Here, +20 dB spectral peaks in the fourth epoch of offset and single conditions produced SCEs, but +10 dB amplification was insufficient for producing SCEs. This is despite the fact that +10 dB and even +5 dB spectral peaks added to sentence contexts have produced SCEs in categorization of these same vowel stimuli (Stilp *et al.*, 2015; n.b., spectral peaks in speech contexts occurred intermittently throughout its entire duration and not just in the last 500 ms). Third, SCE magnitudes affecting categorization of the same vowel stimuli were considerably smaller following speech-shaped noise contexts compared to speech contexts in other studies. As speech studies amplified key frequency regions (100–400 Hz or 550–850 Hz) throughout the context stimuli, the closest comparison would be to epoch 4 in offset and onset paradigms where the spectral peak persisted throughout the context. Amplification of key frequency regions in speech contexts by +10 dB produced SCE magnitudes of 1.24 stimulus steps (Stilp *et al.*, 2015) but only 0.41–0.43 steps here (Figs. 2 and 3). Likewise, amplifying these frequency regions by +20 dB in speech contexts produced SCE magnitudes of 1.54–1.77 stimulus steps (Stilp *et al.*, 2015; Assgari and Stilp, 2015) compared to 0.62–0.76 stimulus steps here (Figs. 2 and 3). These differences in absolute effect magnitudes likely reflect the stark acoustic differences between sentences and sentence-length speech-shaped noise. Watkins' (1991) investigation of different contexts producing SCEs in vowel categorization produced a similar finding: sentence contexts produced SCEs nearly twice as large as those observed following noise contexts (which were filtered throughout their entire duration, akin to epoch 4 of offset and onset paradigms here). Additionally, spectral peaks were presented in specific temporal positions of the noise context and with fixed durations. This is notably different from speech, where acoustic energy waxes and wanes as functions of speaking rate and phonemic content. These modulations are tied to the syllabic structure of speech, which produces temporal oscillations at a rate of roughly 2–8 Hz (Houtgast and Steeneken, 1985). In the present study, context spectral peak durations being multiples of 500 ms broadly resembles peaks that would be produced by a slow speaking rate of two syllables per second (2 Hz). However, energy in the F_1 frequency regions of the noise (as well as neighboring frequency regions) was constant, unlike the comparatively spectrally sparse signal that is speech.

Physiological and/or psychoacoustic considerations of the noise contexts might also illuminate these discrepancies in results. Masking of spectral peaks in the context might underlie smaller SCE magnitudes overall in noise-context studies. In studies employing speech contexts, acoustic energy waxes and wanes over time and over frequency.

Here, the noise context had constant energy at frequencies flanking the spectral peak that produced the SCE. This constant energy likely (at least partially) masked the spectral peak through the upward spread of masking. This "self-masking" would decrease the efficacy with which the spectral peak produced the subsequent SCEs, analogous to decreasing filter gain (which reliable decreases the magnitude of the resulting SCE; Stilp *et al.*, 2015; Stilp and Assgari, 2017; Frazier *et al.*, 2019). Additionally, one cannot rule out contributions of forward masking when spectrally neutral context preceded the subsequent spectral peak (offset epochs 1–3, single epochs 2–4). Either or both of these sources of masking might explain why SCEs were extinguished altogether for +10 dB peaks limited to the last 500-ms epoch (offset epoch 1; single epoch 4).

Differences in results across noise-context and speech-context studies might also be due to higher-level contributions to SCEs. While nonspeech contexts clearly can influence speech sound categorization through SCEs [e.g., Watkins (1991), Lotto and Kluender (1998), and Holt (2005)], other factors might affect the relative magnitudes of these effects. First, noise contexts and speech targets are discontinuous sounds, clearly emanating from different sources. This discontinuity may hinder listeners' ability to perceptually group these sounds together on a given trial as belonging to the same auditory stream. This challenge to sequential grouping might reduce the impact of the context effect on categorization of the target stimulus. Second, the present stimuli might also present a case of the "old-plus-new heuristic" of Bregman (1990), where an ongoing sound that suddenly becomes more intense and/or more complex can be interpreted as a new sound joining an old sound. In this case, the context becoming more complex through the addition of a spectral peak partway through the trial might be interpreted as a new sound joining an old sound rather than a single sound with frequencies being (suddenly) enhanced by the listening environment. This would be consistent with smaller SCEs being produced by the noise contexts tested here compared to speech contexts,[5] and might also explain how a small spectral peak in the last epoch (the "new" sound) was ineffective in altering performance given that it followed 1500-ms of spectrally neutral noise (the "old" sound; cf. +10 dB conditions of offset epoch 1 and single epoch 4). As reviewed in Sec. I, SCEs are not exclusively peripheral phenomena but also receive contributions from central processing (Watkins, 1991; Holt and Lotto, 2002; Feng and Oxenham, 2018; Stilp, 2020b) and attention (Feng and Oxenham, 2018; Bosker *et al.*, 2020). Future research that delineates the relative contributions of lower-level processing and higher-level factors to SCEs (and acoustic context effects more broadly) will be highly illuminating.

Natural sounds in the acoustic environment are highly variable, as has been widely documented for speech sounds (i.e., the lack of invariance). Acoustic context effects (including SCEs) provide some resilience to this problem, as extreme acoustic variability within sounds is alleviated when perception is magnifying changes in acoustic properties between sounds (Stilp, 2020a). Noise was utilized as a context stimulus here to control spectrotemporal variability in the effort to elucidate key temporal characteristics of SCEs in speech sound categorization. Some patterns of results were well-predicted by previous findings, such as longer-duration and/or higher-magnitude spectral peaks in the context producing larger SCEs. Other results were contrary to predictions, such as the target categorization being biased by a spectral peak at the beginning of the context that was followed by up to 1500 ms of spectrally neutral context. This complex pattern of results might reflect the multiple contributions to SCEs in the auditory system, from peripheral and central neural processing to various higher-level processes discussed above (including but not limited to attention, grouping, and segregation). For example, Sjerps and colleagues (2019) discussed multiple concurrent types of normalization (contrast enhancement, estimations relative to talker acoustics, and expectations) potentially shaping their recordings of SCEs in human auditory cortex. Also, the specific goals of the experiment merit consideration. The base phenomenon of the SCE has been widely reported in the literature, having been observed for every combination of speech and nonspeech context and target stimuli [see Stilp, (2020a) for review]. However, the mere presence or absence of the effect is a less strenuous test than defining more specific characteristics of the effect, and these finer tests may exhibit a particular sensitivity to certain stimulus properties. Returning to the question of the timecourse of SCEs, results pooled across studies presenting various context stimuli might not all neatly fit onto a single timecourse. Speech sound categorization has been suggested to be shaped by the last 500 ms of speech contexts (Stilp, 2018), the global mean of 2100-ms pure tone contexts (Holt, 2006), and a mixture of later-occurring (offset) and earlier-occurring (onset, single) spectral peaks in 2000-ms noise contexts here. Further research is needed to clarify the reasons for and/or mechanisms underlying these varying estimates. Therefore, it is essential to not only give careful consideration to stimulus selection / construction in these studies, but also to keep results in the context of other findings as these finer characteristics of SCEs (and other acoustic context effects) are pursued.

ACKNOWLEDGMENTS

This project was funded by a grant from the University of Louisville Executive Vice President for Research and Innovation Internal Grant Program. We wish to acknowledge Associate Editor Joshua Bernstein and two anonymous reviewers for their helpful feedback and suggestions. We also thank Samantha Cardenas, Rebecca Davis, Jonathan Frazier, Joshua Lanning, and Caroline Smith for their assistance scheduling and running participants. All data and analysis scripts are available at https://osf.io/n5vym.

[1]Due to an equipment error, all sounds in one block were presented 5.4 dB SPL higher than these levels for two participants. For participant #9 completing the single paradigm with +20 dB spectral peaks, context presentation levels were 78 dB SPL instead of 73 dB SPL. For participant #10 completing the onset paradigm with +10 dB spectral peaks, context presentation levels ranged from 71–76 dB SPL instead of 66–71 dB SPL. This did not have any systematic effect on their performance, as adjudged by comparing their SCE magnitudes in the affected conditions to those of the rest of the sample.

[2]Epoch 4 was chosen as the reference level for two reasons. First, proximal effects of preceding spectral context are abundant in the speech perception literature [see Stilp (2020a) for review], which makes epoch 4 most likely to elicit an SCE in all paradigms. Second, it was the epoch expected to produce the largest SCEs in all three experimental paradigms.

[3]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0006657 for the full results of the mixed-effects models comparing performance in individual epochs across offset, onset, and single paradigms.

[4]The 500 ms epoch duration tested here exceeds the time constant of adaptation in the auditory periphery [e.g., Westerman and Smith (1984)]. It is unlikely that peripheral adaptation was the sole mechanism contributing to the present results, but nevertheless peripheral processing does appear to contribute more to SCEs than central processing (Watkins, 1991; Holt and Lotto, 2002; Feng and Oxenham, 2018; Stilp, 2020b).

[5]Speech contexts do not always produce larger SCEs than nonspeech contexts. Contexts comprised of sequences of pure tones have been reported to produce equivalent or even larger SCEs than speech contexts (Huang and Holt, 2012; Laing *et al*., 2012).

ANSI (**2010**). *S3.6-2010: American National Standards Specifications for Audiometers* (American National Standards Institute, New York).

Assgari, A. A., and Stilp, C. E. (**2015**). "Talker information influences spectral contrast effects in speech categorization," J. Acoust. Soc. Am. **138**(5), 3023–3032.

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," http://cran.r-project.org/package=lme4 (Last viewed 10/4/2021).

Boersma, P., and Weenink, D. (**2014**). "Praat: Doing phonetics by computer [computer program]."

Bosker, H. R., Sjerps, M. J., and Reinisch, E. (**2020**). "Spectral contrast effects are modulated by selective attention in 'cocktail party' settings," Atten. Percept. Psychophys. **82**, 1318–1332.

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

Broadbent, D. E., and Ladefoged, P. (**1960**). "Vowel judgements and adaptation level," Proc. R. Soc. B: Biol. Sci. **151**, 384–399.

Delgutte, B. (**1996**). "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell, Oxford), pp. 507–538.

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

Feng, L., and Oxenham, A. J. (**2018**). "Spectral contrast effects produced by competing speech contexts," J. Exp. Psych. Human Percept. Perform. **44**(9), 1447–1457.

Frazier, J. M., Assgari, A. A., and Stilp, C. E. (**2019**). "Musical instrument categorization is highly sensitive to spectral properties of earlier sounds," Atten. Percept. Psychophys. **81**(4), 1119–1126.

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

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.

Holt, L. L., and Lotto, A. J. (**2002**). "Behavioral examinations of the level of auditory processing of speech context effects," Hear. Res. **167**(1–2), 156–169.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech-intelligibility in auditoria," J. Acoust. Soc. Am. **77**(3), 1069–1077.

Huang, J., and Holt, L. L. (**2012**). "Listening for the norm: Adaptive coding in speech categorization," Front. Psych. **3**, 1–6.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**(1), 98–104.

Laing, E. J., Liu, R., Lotto, A. J., and Holt, L. L. (**2012**). "Tuned with a tune: Talker normalization via general auditory processes," Front. Psych. **3**, 1–9.

Lindblom, B. E., and Studdert-Kennedy, M. (**1967**). "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Am. **42**(4), 830–843.

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.

Mann, V. A. (**1980**). "Influence of preceding liquid on stop-consonant perception," Percept. Psychophys. **28**(5), 407–412.

Mann, V. A., and Repp, B. H. (**1981**). "Influence of preceding fricative on stop consonant perception," J. Acoust. Soc. Am. **69**(2), 548–558.

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (**2007**). "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals," J. Speech Lang. Hear. Res. **50**(4), 940–967.

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

Pressnitzer. D., Sayles. M., Micheyl. C., and Winter, I. M. (**2008**). "Perceptual organization of sound begins in the auditory periphery," Curr. Bio. **18**(15), 1124–1128.

R Development Core Team (**2021**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.org/ (Last viewed 10/4/2021).

Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (**2019**). "Speaker-normalized sound representations in the human auditory cortex," Nat. Commun. **10**, 1–9.

Sjerps, M. J., Zhang, C., and Peng, G. (**2018**). "Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context," J. Exp. Psych. Human Percept. Perform. **44**(6), 914–924.

Stilp, C. E. (**2018**). "Short-term, not long-term, average spectra of preceding sentences bias consonant categorization," J. Acoust. Soc. Am. **144**(3), 1797.

Stilp, C. E. (**2020a**). "Acoustic context effects in speech perception," Wiley Interdisc. Rev. Cogn. Sci. **11**(1–2), 1–18.

Stilp, C. E. (**2020b**). "Evaluating peripheral versus central contributions to spectral context effects in speech perception," Hear. Res. **392**, 107983–107912.

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

Stilp, C. E., Anderson, P. W., and Winn, M. B. (**2015**). "Predicting contrast effects following reliable spectral properties in speech perception," J. Acoust. Soc. Am. **137**(6), 3466–3476.

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 Winn, M. B. (**2021**). "The timecourse and neural mechanisms of the influence of precursor spectral properties on speech perception," in *44th Annual MidWinter Meeting of The Association for Research in Otolaryngology*.

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

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

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

Westerman, L. A., and Smith, R. L. (**1984**). "Rapid and short-term adaptation in auditory nerve responses," Hear. Res. **15**(3), 249–260.

Winn, M. B., and Litovsky, R. Y. (**2015**). "Using speech sounds to test functional spectral resolution in listeners with cochlear implants," J. Acoust. Soc. Am. **137**(3), 1430–1442.