JASA ARTICLE

Context effects in perception of vowels differentiated by F_1 are not influenced by variability in talkers' mean F_1 or F_3

Hannah E. Mills,[1] Anya E. Shorey,[1] Rachel M. Theodore,[2] and Christian E. Stilp[1,a)]
[1]*Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292, USA*
[2]*Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, Connecticut 06269, USA*

ABSTRACT:
Spectral properties of earlier sounds (context) influence recognition of later sounds (target). Acoustic variability in context stimuli can disrupt this process. When mean fundamental frequencies (f0's) of preceding context sentences were highly variable across trials, shifts in target vowel categorization [due to spectral contrast effects (SCEs)] were smaller than when sentence mean f0's were less variable; when sentences were rearranged to exhibit high or low variability in mean first formant frequencies (F_1) in a given block, SCE magnitudes were equivalent [Assgari, Theodore, and Stilp (2019) J. Acoust. Soc. Am. **145**(3), 1443–1454]. However, since sentences were originally chosen based on variability in mean f0, stimuli underrepresented the extent to which mean F_1 could vary. Here, target vowels (/ɪ/-/ɛ/) were categorized following context sentences that varied substantially in mean F_1 (experiment 1) or mean F_3 (experiment 2) with variability in mean f0 held constant. In experiment 1, SCE magnitudes were equivalent whether context sentences had high or low variability in mean F_1; the same pattern was observed in experiment 2 for new sentences with high or low variability in mean F_3. Variability in some acoustic properties (mean f0) can be more perceptually consequential than others (mean F_1, mean F_3), but these results may be task-dependent. © *2022 Acoustical Society of America*. https://doi.org/10.1121/10.0011920

(Received 8 July 2021; revised 3 May 2022; accepted 8 June 2022; published online 1 July 2022)
[Editor: Benjamin V. Tucker] Pages: 55–66

I. INTRODUCTION

All perception takes place in context. Perception of a sound depends on its acoustic properties and also on the acoustic properties of sounds that precede or follow it in time (i.e., the context). When successive sounds differ in their acoustic properties, this difference can be perceptually magnified via a contrast effect. For example, the vowels /ɪ/ (as in "bit") and /ɛ/ (as in "bet") differ primarily in the frequency of their first formant (F_1). When the vowel is preceded by a context sentence with F_1 frequencies occurring in a lower frequency range (which is typical of /ɪ/), listeners perceive /ɛ/ more often; when the vowel is preceded by a context sentence with F_1 frequencies occurring in a higher frequency range (which is typical of /ɛ/), listeners perceive /ɪ/ more often (Ladefoged and Broadbent, 1957). This is known as a spectral contrast effect (SCE). SCEs are pervasive in speech perception (Stilp, 2020).

Historically, SCEs have been measured with every trial presenting renditions of the same context stimulus that differed only in spectral composition (e.g., F_1 frequencies in a sentence occurring in either lower or higher frequency ranges). Assgari and Stilp (2015) examined whether SCEs were sensitive to different degrees of acoustic variability across the context sentences. All context sentences were filtered to amplify either low-F_1 frequencies (100–400 Hz) or high-F_1 frequencies (550–850 Hz) by +5 dB to produce an SCE in categorization of the test vowels (/ɪ/-/ɛ/). In one condition, filtered renditions of the same context sentence were presented on each trial (consistent with past approaches). In a second condition, 200 different context sentences were presented in the testing block, all spoken by a single talker. In a third condition, 200 different context sentences spoken by 200 unique talkers were presented in the testing block. The magnitudes of SCEs affecting vowel categorization were comparable when context sentences (whether 1 or 200) were spoken by a single talker, but SCEs were diminished when context sentences were spoken by 200 different talkers. Thus, talker variability in the context sentences decreased the magnitudes of SCEs. Diminished SCEs challenge speech perception because a mechanism by which perceptually ambiguous sounds are disambiguated by surrounding context is limited.

A litany of studies have documented processing costs associated with perceiving speech from multiple talkers compared to perceiving speech from a single talker (Creelman, 1957; Assmann *et al.*, 1982; Mullennix *et al.*, 1989; Mullennix and Pisoni, 1990; Goldinger, 1996; Magnuson and Nusbaum, 2007; Zhang and Chen, 2016; Choi *et al.*, 2018; Stilp and Theodore, 2020). While this literature had not previously included examinations of acoustic context effects, the results of Assgari and Stilp (2015) were consistent with this overall pattern. Importantly, some talker adaptation studies suggested that these processing costs are due in part to variability in the talkers' fundamental frequency (f0) characteristics. Magnuson and Nusbaum

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

(2007) had participants listen to words spoken by one talker or two talkers. Participants either heard two male voices differing greatly in f0, a male and female differing greatly in f0, or two men differing slightly in f0. When f0 differed greatly across talkers, participants responded faster to words spoken by one talker than words spoken by multiple talkers. When f0 was similar across talkers, reaction times were similar when hearing one or two talkers. Goldinger (1996) measured participants' word recognition when words were spoken by different talkers. When the difference in f0 between talkers was larger, listeners had worse recall accuracy and slower reaction time compared to voices with a smaller difference in f0. Stilp and Theodore (2020) examined performance in a speeded word recognition task for words spoken by either the same talker or by different talkers. Crucially, mean f0 characteristics of talkers in mixed-talker blocks exhibited either low variability or high variability. Listeners' response times in the mixed-talker blocks increased as f0 variability increased.

Regarding acoustic context effects, in a *post hoc* analysis of their stimuli, Assgari and Stilp (2015) suggested that variability in talkers' f0's might have contributed to the reduction of SCE magnitudes, but this was not explicitly controlled in the experiment. To directly test the perceptual influence of f0 variability on context effects, Assgari *et al.* (2019) constructed two sets of 40 context sentences. Each set was comprised of 20 talkers who were men and 20 talkers who were women, but one set exhibited high variability across the mean f0's of context sentences, while the other set exhibited low variability across mean f0's (Fig. 1, top row, left column). When the mean f0's of context sentences were highly variable from trial to trial, SCEs in vowel categorization were smaller than when the mean f0's of context sentences were less variable (Assgari *et al.*, 2019). This was consistent with other reports of variability in talkers' f0 characteristics challenging speech perception (Goldinger, 1996; Magnuson and Nusbaum, 2007; Stilp and Theodore, 2020).

Different talkers' voices vary in myriad acoustic properties. While sentences in Assgari *et al.* (2019) were presented based on low or high variability in mean f0, talkers' voices were concurrently varying in many other acoustic properties beyond f0. For example, the second and third columns in the top row of Fig. 1 illustrate variability in mean F_1 and mean F_3 that was not controlled in the experiment of Assgari *et al.* (2019). While they attributed their results to variability in mean f0, contributions of other sources of acoustic variability to the results were unclear. Assgari *et al.* (2019) examined the perceptual influence of variability in the mean F_1 frequencies of context sentences, as this was predicted to be highly relevant to the categorization of target vowels that differed principally in F_1. In a follow-up experiment, context sentences that were selected to have low variability or high variability across their mean f0's (their experiment 2) were regrouped to form conditions that exhibited low variability or high variability in mean F_1 frequencies (their experiment 3; second row of Fig. 1). Variability in mean f0 was equated across the two blocks, so it would not differentially affect SCE magnitudes in the presence of low/high variability in mean F_1. SCE magnitudes did not vary as a function of variability in mean F_1, leading Assgari *et al.* (2019) to conclude that variability in the mean F_1 frequencies of context sentences did not have the same detrimental effect on SCEs as did variability in mean f0's. However, the stimuli in this follow-up experiment were originally selected based on their measures of (and variability in) f0, not F_1. This was not representative of how much F_1 can vary across talkers (cf. the central panel of the second row of Fig. 1). As such, this finding might not be the best gauge of whether and how variability in sentences' mean F_1 frequencies can influence SCEs. Experiment 1 of the present report provides a more sensitive test of this question. Context sentences were selected and presented based on their measures of mean F_1, again testing whether variability in this acoustic property modulates SCE magnitudes in categorization of the same /ɪ/-/ɛ/ target vowels.

F_1 frequencies may differ across talkers as a function of vocal tract length, but it can be argued that its principal role in speech is indicating phonetic identity (e.g., vowel height). While variability in mean F_1 might be detrimental for categorizing target vowels that are cued primarily by F_1, this variability might be inconsequential to perception owing to its habitual variability via signaling different phonemes. Alternatively, higher formants have been proposed to play a role in cueing a talker's vocal tract length (Johnson and Sjerps, 2021). For example, F_3 frequency has been used to estimate the talker's vocal tract length (Nordström and Lindblom, 1975) and may serve as a perceptual anchor whose relation to lower vowel formants resolves some of the high acoustic variability across different talkers' productions (Peterson, 1951). Higher formants are less sensitive to speech articulation than lower formants (Wakita, 1977; Lammert and Narayanan, 2015), so variability in this property of talkers' speech may prove more consequential to talker perception than variability in lower formants. To the extent that F_3 more closely reflects talker-specific characteristics than F_1 does, variability in this acoustic feature would be expected to diminish context effects in vowel categorization in the same way that f0 variability does (Assgari *et al.*, 2019). Experiment 2 selected and presented context sentences based on their measures of mean F_3, otherwise utilizing the same paradigm as experiment 1 (Fig. 1, bottom row).

According to the efficient coding hypothesis (Attneave, 1954; Barlow, 1961), sensory and perceptual systems capitalize on structure in the sensory environment, as that makes neural and/or perceptual processing efficient. The efficient coding hypothesis has a long and rich history in its applications to neural processing and perception in the visual system. A wide range of studies have documented the statistical structure of natural images (Field, 1987; Olshausen and Field, 1996). This stimulus structure has been linked to neural response properties in the visual system (Field, 1987; Ruderman *et al.*, 1998; Schwartz and Simoncelli, 2001) and performance in visual perception tasks (Geisler *et al.*, 2001;

https://doi.org/10.1121/10.0011920

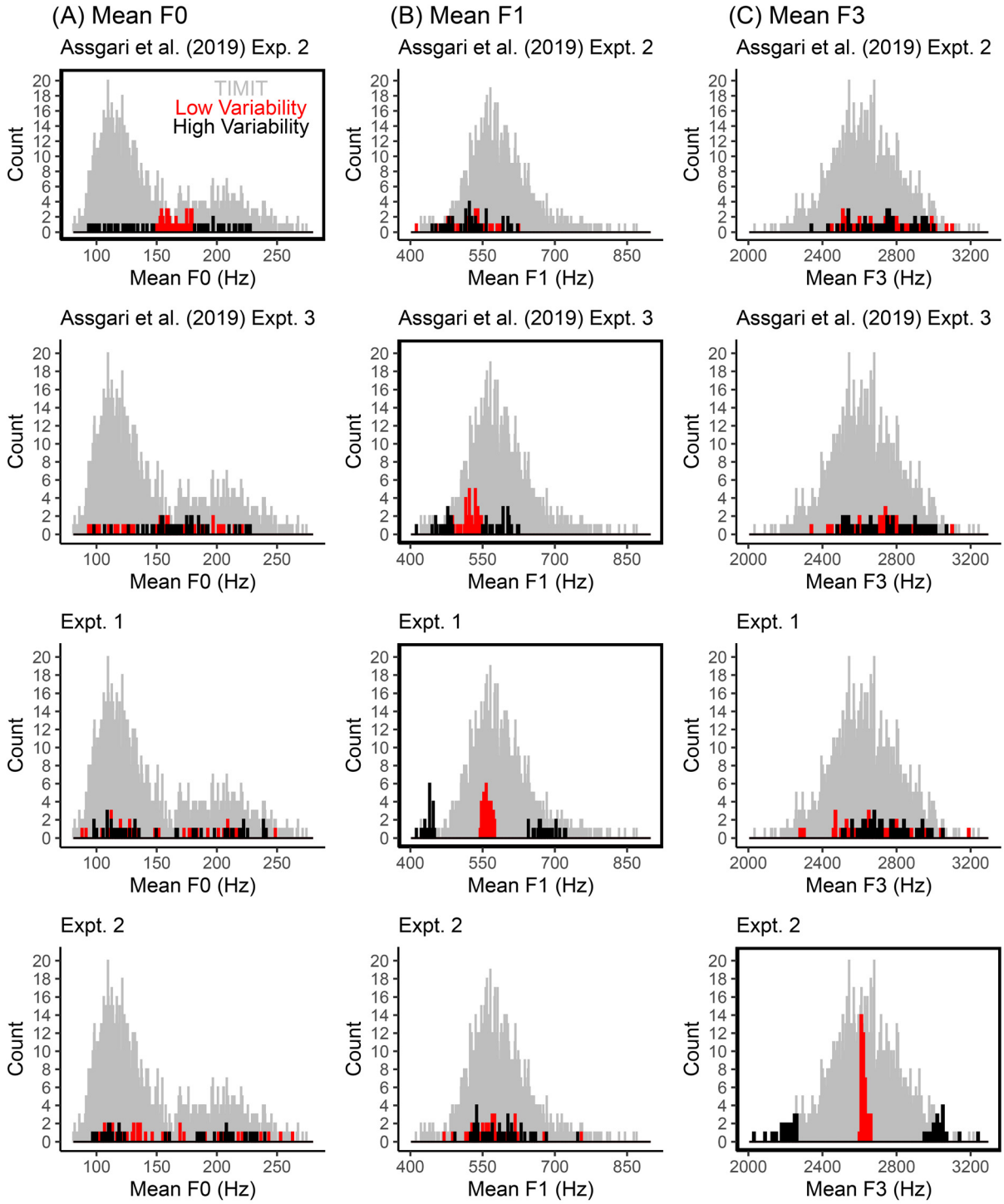

FIG. 1. (Color online) (A) Histograms of mean f0 (left column), mean F_1 (middle column), and mean F_3 (right column) measurements in sentences from the TIMIT database. Each row depicts stimuli tested in a single experiment: experiment 2 of Assgari *et al.* (2019) (first row); experiment 3 of Assgari *et al.* (2019) (second row); and in the present report, experiment 1 (third row); and experiment 2 (fourth row). In each histogram, the low variability condition is depicted in red, and the high variability condition is depicted in black. For comparison, each panel also shows the mean frequencies for 2019 TIMIT sentences plotted in gray using a thinner bin width. Panels with boxes around them indicate experiments that were designed to test perceptual sensitivity to low or high variability in that particular metric.

Tkačik *et al.*, 2010). While applications of the efficient coding hypothesis to auditory perception (and more specifically to speech perception) are comparatively nascent, they show similar promise and productivity (Kluender *et al.*, 2013; Kluender *et al.*, 2019; Gervain and Geffen, 2019). For instance, considerable covariance is shared among talkers' fundamental and formant frequencies in vowel production (Kluender *et al.*, 2013), and vowel identification suffers when these natural

correlations are violated (Assmann and Nearey, 2008). Additionally, speeded word identification was fastest when stimuli were spoken by a single talker (i.e., highly structured), slower when words were spoken by acoustically similar (in terms of mean f0) talkers (i.e., intermediate level of structure), and significantly slowed again when words were spoken by acoustically dissimilar talkers (i.e., low structure; Stilp and Theodore, 2020). In studies examining the influence of talker variability on SCEs (Assgari and Stilp, 2015; Assgari *et al.*, 2019), conditions featuring low acoustic (f0) variability indicated a degree of structure in the acoustic environment, and SCEs were facilitated by maintaining their magnitudes (particularly in comparison to hearing the same talker on every trial; Assgari and Stilp, 2015). Conditions featuring high acoustic (f0) variability indicated less structure in the environment, which challenged perception and diminished SCE magnitudes. In the present study, vowel categorization is predicted to be facilitated by the presence of structure (low acoustic variability across talkers' voices) and challenged by the relative lack of structure (high variability). Thus, SCE magnitudes are predicted to be smaller when context sentences exhibit high variability in mean F_1 as compared to low variability in mean F_1 (experiment 1); similarly, sentences with high variability in mean F_3 are predicted to produce smaller SCEs than sentences with low variability in mean F_3 (experiment 2). Should results contradict these predictions, it may indicate that perception does not leverage any and all structure in the input, but that structure in some acoustic properties (mean f0) influences perception more than structure in other properties (mean F_1 and/or mean F_3). Such results would necessitate refinement to efficient coding approaches to speech perception.

II. EXPERIMENT 1

A. Methods

1. Participants

Forty-eight undergraduate students at the University of Louisville participated in exchange for course credit. All participants reported normal hearing and were native English speakers. From this sample, 40 listeners completed the experiment, and responses from 20 of these listeners were included in data analyses. As detailed below, reasons for exclusion from analyses included failing a headphone screener ($n = 6$), inability to reliably distinguish endpoints of the vowel continuum during a practice session ($n = 6$), or inability to maintain that level of accuracy on vowel endpoints throughout the main experiment ($n = 8$).

2. Stimuli

a. Context sentences. Sentence selection followed a similar process as that reported in Assgari *et al.* (2019). All sentences in TIMIT (Garofolo *et al.*, 1990) dialect regions DR3 (North Midland) and DR4 (South Midland) were analyzed by a custom Praat (Boersma and Weenink, 2019) script that used linear predictive coding to estimate f0 contours and formant contours. f0, F_1, and F_3 contours were visually inspected, and any aberrant estimates in pitch/formant tracking were manually removed. Then mean f0, mean F_1, and mean F_3 values were calculated for each sentence (cf. Fig. 1). From these analyses, 40 sentences were selected for presentation in the low variability in mean F_1 condition, and a separate set of 40 sentences were selected for presentation in the high variability in mean F_1 condition (for more details, see supplementary Table I).[1] Each condition contained 20 talkers who were men and 20 talkers who were women, and no talker was presented in both conditions. Stimulus sets were carefully constructed to differ markedly in terms of F_1 variability [standard deviation (SD) of mean F_1 in low variability sentences = 8.25; SD of mean F_1 for high variability sentences = 124.42] but be well-matched in terms of the grand means of mean F_1 (low variability: grand mean = 559.63; high variability: grand mean = 559.08) and mean f0 (low variability: grand mean = 162.62 Hz, SD = 47.62; high variability: grand mean = 160.17 Hz, SD = 48.74). F_3 measures were not considered for this experiment, resulting in a slightly lower grand mean of mean F_3 with slightly higher variability in the low variability block (mean = 2670.46 Hz, SD = 185.55) compared to high variability block (mean = 2746.19 Hz, SD = 135.30).

b. Vowel targets. Target vowels were the same ten-step continuum ranging from /ɪ/ to /ɛ/ as tested in previous investigations of SCEs (Assgari and Stilp, 2015; Assgari *et al.*, 2019). Briefly, vowels were synthesized based on natural recordings from a talker who was a man. These speech samples were resynthesized using linear predictive coding in Praat. The /ɪ/ endpoint had an F_1 that linearly increased from 400 to 430 Hz, while F_2 linearly decreased from 2000 to 1800 Hz. The /ɛ/ endpoint had an F_1 that linearly decreased from 580 to 550 Hz, while F_2 linearly decreased from 1800 to 1700 Hz. The vowel continuum was created by taking these endpoint vowels and linearly morphing their formant tracks through a script in Praat (Winn and Litovsky, 2015). Final vowel stimuli were 246 ms in duration with an f0 set to 100 Hz throughout the vowel.

All filtered sentences and target vowels were equated in root mean square amplitude. Experimental trials consisted of a filtered sentence followed by a 50-ms silent interstimulus interval and then a target vowel. All TIMIT sentences were upsampled from their native sampling rate of 16 000 Hz to 44 100 Hz, matching the sampling rate of the vowel targets.

3. Procedure

The experiment was administered online using the Gorilla testing platform (Anwyl-Irvine *et al.*, 2020). Participants were sent a link to the experiment and completed it on a personal computer outside of the laboratory. To standardize sound presentation, participants first completed a screener to confirm they were wearing headphones (Woods *et al.*, 2017). On each of six trials, listeners heard

https://doi.org/10.1121/10.0011920

three tones and were asked to report which was the quietest. The correct answer was the tone that was −6 dB relative to the two other tones, but the foil answer was the tone that was presented 180° out of phase across the stereo channels. Listening over headphones promotes selection of the correct answer (−6 dB tone); listening over speakers promotes selection of the foil answer (180° out of phase tone, which is quieter over speakers due to destructive interference). Participants were required to identify the correct answer on five out of six trials. If they did not meet this criterion, then they were allowed to repeat the headphone screen one additional time. Six participants did not meet this criterion on either headphone screen; their responses were removed from statistical analyses.

Second, listeners completed a set of 20 practice trials. Each trial presented a sentence from the AzBio corpus (Spahr *et al.*, 2012) followed by one of the two continuum endpoint vowels (as categorizing endpoints of the vowel continuum is objectively correct or incorrect). The listener pressed the "i" key to label the target vowel as "ih" as in "bit" or pressed the "e" key to label the target vowel as "eh" as in "bet." If the listener failed to reach 80% accuracy on endpoint vowels after one block, practice trials were repeated up to two more times to reach 80% accuracy. If after three blocks of practice trials the listener did not achieve 80% accuracy, their results were removed from subsequent statistical analyses. Six participants did not achieve 80% accuracy during practice.

Next, the main experiment consisted of two blocks (low variability in mean F_1, high variability in mean F_1). Each block consisted of 160 trials (four repetitions of each unique sentence) and took about 12 min to complete. Block order was counterbalanced across participants. Participants were allowed to take breaks in between blocks. The entire session lasted approximately 40 min.

B. Results

As noted above, of the 40 participants who completed the experiment, 12 failed at least one of the two screeners (headphones screen, practice block). An additional performance criterion was implemented of maintaining at least 80% accuracy on vowel continuum endpoints throughout the main experiment. Eight participants failed to meet this performance criterion, so their responses were not included in data analyses. This resulted in the final sample size of 20 listeners, matching Assgari *et al.* (2019).

Trial-level data were analyzed in a generalized linear mixed-effects model in R (R Development Core Team, 2021) using the lme4 package (Bates *et al.*, 2014) with the binomial logit linking function. The dependent measure was vowel identification (/ɪ/ = 0, /ɛ/ = 1). The aim was to use the same fixed and random effects structure as that tested in experiments 2 and 3 of Assgari *et al.* (2019): fixed effects of target, filter, variability, and all interactions between these factors; random intercepts by subject; and random slopes by subject for target, filter, and variability. However, the model with this architecture analyzing the present data did not converge. Stepwise modeling revealed that the maximal random effects structure that facilitated model convergence was random intercepts for subjects and random slopes for target; all fixed effects were retained as described above. Target was entered into the model as a continuous variable (steps 1–10, centered around the mean). Sum coding was used for the fixed effects of filter (high F_1 amplification = −0.5, low F_1 amplification = +0.5) and variability (low variability = −0.5, high variability = +0.5).

Model results are listed in Table I and visualized in Fig. 2 in terms of /ɛ/ responses as functions of the fixed effects, as created using the interactions package in R (Long, 2019). As expected, the model reports a significant effect of target, such that each rightward step along the vowel continuum (toward higher F_1 values and the /ɛ/ endpoint) increased the log odds of participants responding /ɛ/. There was a main effect of context filter, indicating that changing the filtering condition from a high-F_1-amplified context to a low-F_1-amplified context increased the probability of /ɛ/ responses, confirming the presence of SCEs. There was a significant main effect of variability, such that listeners responded /ɛ/ more often when hearing high variability in mean F_1 sentences than when hearing low variability in mean F_1 sentences. The interaction between target and filter was also significant, indicating that the slope of the psychometric function was steeper in the low-F_1-amplified condition. The key analysis of interest, the interaction between filter and variability, was not statistically significant. Put another way, the magnitudes of SCEs were comparable following low variability in mean F_1 sentences and high variability in mean F_1 sentences.

Next, a mixed-effects model analysis was conducted to quantitatively compare patterns of results across the present experiment and experiment 3 of Assgari *et al.* (2019). Responses were coded as before (/ɪ/ = 0, /ɛ/ = 1), and all main effects and interactions between target, filter, variability, and experiment were included. Filter (high-F_1-amplified = −0.5, low-F_1-amplified = +0.5) and variability (low = −0.5, high = +0.5) were sum-coded as above, and experiment was sum-coded as well (the present experiment = −0.5, 2019

TABLE I. Beta estimate ($\hat{\beta}$), standard error (SE), *Z*-statistic, and *p*-value 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. Filter and variability were sum-coded; the level associated with the −0.5 contrast for each factor is shown in parentheses.

	$\hat{\beta}$	SE	*Z*	*p*
Intercept	0.213	0.140	1.525	0.127
Target	0.905	0.072	12.628	<0.001
Filter (high F_1)	0.301	0.074	4.048	<0.001
Variability (low)	0.351	0.074	4.716	<0.001
Target × filter	0.103	0.036	2.893	0.004
Target × variability	0.016	0.036	0.436	0.663
Filter × variability	0.026	0.149	0.174	0.862
Target × filter × variability	−0.021	0.071	−0.300	0.764

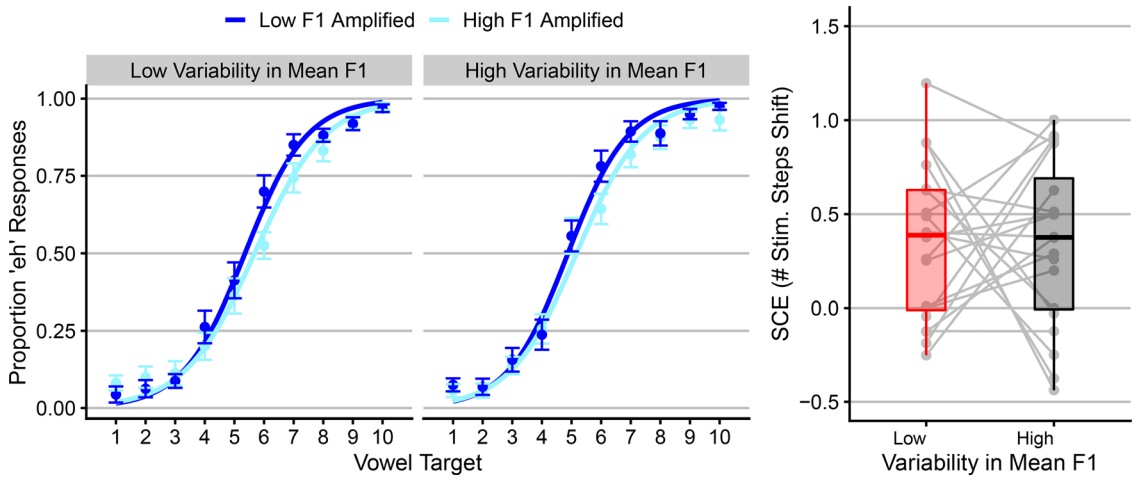

FIG. 2. (Color online) (Left) The mixed-effects model fit to listeners' responses as a function of vowel target in experiment 1. Dark blue lines represent responses following low-F_1-amplified context sentences, light blue lines represent responses following high-F_1-amplified context sentences. Circles depict mean proportions of "eh" responses; error bars depict one standard error of the mean. (Right) SCE magnitudes (the number of stimulus steps separating 50% points on the psychometric functions) calculated for each listener in each variability condition (red indicating low variability in mean F_1, black indicating high variability in mean F_1, consistent with coloration of Fig. 1). Each gray line connects one listener's SCEs in both variability conditions.

experiment = +0.5). A model with random slopes for each fixed main effect and random intercepts for subjects did not converge, so random slopes were added iteratively and retained if the model converged and fit significantly improved. The final model had random slopes for target and variability as well as random intercepts for subjects. The interaction between filter and experiment was significant ($Z = 2.043$, $p = 0.041$), indicating that SCE magnitudes were smaller in the present experiment compared to the previous experiment. Critically, the three-way interaction between filter, variability, and experiment was not significant ($Z = 0.018$, $p = 0.985$), indicating that the null difference in SCE magnitudes observed across low variability in mean F_1 and high variability in mean F_1 conditions in Assgari *et al.* (2019) was replicated here. The full model results are available in supplementary Table II.[1]

The present results were also tested against those from experiment 2 of Assgari *et al.* (2019) to establish whether SCE magnitudes patterned differently across wide ranges of variability in mean f0 and mean F_1. The model had the same fixed effects structure as detailed above, but the maximal random effects structure that converged included random slopes for target and variability as well as random intercepts for subjects. The three-way interaction between filter, variability, and experiment was significant ($Z = -2.138$, $p = 0.033$), indicating that the relationship between SCEs and acoustic variability differed across experiments. In Assgari *et al.* (2019), SCEs were significantly larger in low variability in mean f0 condition than the high variability in mean f0 condition; here, SCEs were of similar magnitudes across low variability in mean F_1 and high variability in mean F_1 conditions (see Table I). This finding is further supported by the filter × variability interaction being only marginally significant when averaging across these experiments ($Z = -1.927$, $p = 0.054$). The full model results are available in supplementary Table III.[1]

C. Discussion

Variability in the mean F_1 frequency of context sentences did not alter SCE magnitudes in vowel categorization. This result echoes the findings of experiment 3 in Assgari *et al.* (2019). While the present study better represented the full range of values that sentence mean F_1 can take (Fig. 1, second and third row of the center column), it arrived at the same conclusion. This coincides with the suggestion by Assgari *et al.* (2019) that variability in mean f0 and variability in mean F_1 might have different consequences for perception, at least in terms of context effect magnitudes in vowel categorization.

Explicit comparisons between experiments 2 and 3 of Assgari *et al.* (2019) and experiment 1 here elucidate how acoustic variability influences the magnitude of SCEs. SCEs occurred in all three studies, consistent with amplifying low-F_1 frequencies versus high-F_1 frequencies in context sentences. One means of differentiating these studies is by how SCE magnitudes varied as a function of low versus high acoustic variability in the context sentences. In Assgari *et al.* (2019) experiment 2, the significant filter × variability interaction indicated that SCEs were smaller in the high variability in mean f0 condition than in the low variability in mean f0 condition. However, this filter × variability interaction was not significant in their experiment 3 or in the present experiment. Thus, SCEs differed due to variability in mean f0 in the former case, but not variability in mean F_1 as in the latter two cases.

The different perceptual consequences of variability in mean f0 versus variability in mean F_1 are also evident in the target × variability interactions in mixed-effect model analyses. In Assgari *et al.* (2019) experiment 2, the significant target × variability interaction indicated that the task was more difficult in the high variability in mean f0 condition (i.e., the shallower slope of the psychometric function) than the low variability in mean f0 condition (i.e., steeper slope). However, this

https://doi.org/10.1121/10.0011920

interaction was not statistically significant in their experiment 3 nor the present experiment, indicating that variability in mean F_1 did not alter the slope of the psychometric function; both the low variability in mean F_1 and high variability in mean F_1 conditions were equally difficult. This was supported by the three-way interaction among talker, variability, and experiment between the present experiment and experiment 2 of Assgari *et al.* (2019) being statistically significant ($Z = -3.719$, $p = 0.0002$; supplementary Table III),[1] but this same interaction was not statistically significant when comparing the present experiment to experiment 3 of Assgari *et al.* (2019) ($Z = -0.133$, $p = 0.894$; supplementary Table II).[1] Thus, it is not only SCEs that illustrate the differential effects that variability in mean f0 and mean F_1 have on vowel categorization, but the psychometric function slopes also provide evidence for this difference.

The results of experiment 1 are contrary to predictions made by the efficient coding hypothesis, that perception would be facilitated in the presence of structure in the context sentences and challenged by the comparative lack of structure. Despite there being very different amounts of variability in mean F_1 frequencies across the two testing blocks, SCE magnitudes were unaffected. The implications of this point are considered further in Sec. IV.

Mean f0 and mean F_1 are far from the only acoustic characteristics that vary across different talkers' voices. As detailed in the Introduction, F_1 and F_3 vary to different degrees in speech (Lammert and Narayanan, 2015) and may contribute differently to perception of speech sounds versus perception of talker characteristics. To this end, experiment 2 used the same paradigm as experiment 1 but with new sentences arranged into low variability in mean F_3 and high variability in mean F_3 conditions.

III. EXPERIMENT 2

A. Methods

1. Participants

Fifty-eight undergraduate students at the University of Louisville participated in exchange for course credit. All participants reported normal hearing and were native English speakers. None participated in experiment 1. From this sample, 45 listeners completed the experiment, and responses from 19 of these listeners were included in data analyses. As detailed below, reasons for exclusion from analyses included failing a headphone screener ($n = 12$), inability to reliably distinguish endpoints of the vowel continuum during a practice session ($n = 7$), or inability to maintain that level of accuracy on vowel endpoints throughout the main experiment ($n = 7$).

2. Stimuli

a. Context sentences. Sentence selection followed a similar process as experiment 1, but the primary metric of interest was the mean frequency of F_3. All TIMIT sentences in the North Midland and South Midland dialect regions were sorted by their mean F_3 frequencies (Fig. 1, right column). All sentences that were tested in previous experiments [experiment 1 here and experiments 2 and 3 in Assgari *et al.* (2019)] were removed, ensuring that novel sentences were presented in experiment 2. From these remaining sentences, 40 were selected as stimuli in the low variability in mean F_3 condition, and 40 different sentences were selected as stimuli in the high variability in mean F_3 condition (for more details, see supplementary Table IV).[1] Each set of 40 sentences contained 20 men and 20 women talkers, and no talker was presented in both conditions. Stimulus sets were constructed to differ markedly in terms of variability of mean F_3 (SD of mean F_3 across low variability sentences = 14.02; SD of mean F_3 across high variability sentences = 429.16) but be matched in terms of the grand mean of mean F_3 measures (grand mean of mean F_3 across low variability sentences = 2621.91 Hz; grand mean of mean F_3 across high variability sentences = 2614.62 Hz). Stimulus sets were well-matched in terms of mean F_1 characteristics (low variability: grand mean = 580.43 Hz, SD = 50.95; high variability: grand mean = 586.01 Hz, SD = 51.24) and mean f0 characteristics (low variability: grand mean = 162.16 Hz, SD = 47.85; high variability: grand mean = 162.72 Hz, SD = 50.97).

b. Vowel targets. Target vowels were the same as those tested in experiment 1. All filtered sentences and target vowels were again equated in root mean square amplitude, concatenated with a 50-ms silent interstimulus interval, and upsampled to 44 100 Hz.

3. Procedure

The procedure was identical to that of experiment 1. Twelve participants did not meet the performance criterion on the headphone screen portion, and seven participants did not achieve 80% accuracy during practice; their results were removed from subsequent statistical analyses. The main experiment consisted of two blocks (context sentences with low variability in mean F_3, context sentences with high variability in mean F_3). Each block consisted of 160 trials (four repetitions of each unique sentence) and took about 12 min to complete. Block order was counterbalanced across participants. Participants were allowed to take breaks in between blocks. The entire session lasted approximately 40 min.

B. Results

As noted above, of the 45 participants who completed experiment 2, 19 failed at least one of the two screeners (headphones screen, practice blocks). An additional seven participants failed to maintain at least 80% accuracy on vowel continuum endpoints throughout the main experiment. Their responses were not included in data analyses, resulting in the final sample size of 19 listeners.

Trial-level data were analyzed in a generalized linear mixed-effects model in R (R Development Core Team, 2021) using the lme4 package (Bates *et al.*, 2014) with the binomial

TABLE II. $\hat{\beta}$, SE, Z-statistic, and p-value for the fixed effects of the mixed-effects model analyzing experiment 2. As described in the main text, target was entered in the model as a continuous factor, centered around the mean. Filter and variability were sum-coded; the level associated with the −0.5 contrast for each factor is shown in parentheses.

	$\hat{\beta}$	SE	Z	p
Intercept	0.336	0.122	2.750	0.006
Target	0.899	0.062	14.594	<0.001
Filter (high F_1)	0.405	0.077	5.257	<0.001
Variability (low)	−0.102	0.077	−1.330	0.184
Target × filter	0.014	0.038	0.360	0.719
Target × variability	−0.044	0.038	−1.168	0.243
Filter × variability	0.062	0.154	0.402	0.688
Target × filter × variability	0.059	0.075	0.781	0.435

logit linking function. The dependent measure was vowel identification (/ɪ/ = 0, /ɛ/ = 1). The same fixed effects structure as experiment 1 was included: fixed effects of target, filter, variability, and all interactions between these factors. The maximal random effects structure that facilitated model convergence included random intercepts by subject and random slopes by subject for target. Fixed effects were coded in the same manner as described for experiment 1.

Model results are listed in Table II and visualized in Fig. 3 in terms of /ɛ/ responses as predicted by the fixed effects using the interactions package in R (Long, 2019). The model reports expected significant effects of target and of filter, the latter of which again confirmed the presence of SCEs. No other main effects or interactions were significant, suggesting that variability in mean F_3 did not modulate the magnitudes of SCEs (i.e., nonsignificant filter × variability interaction).

Results were analyzed across experiments 1 and 2 in a mixed-effects model analysis. The maximal model that converged included fixed effects of target (mean-centered as described above), filter (high F_1 amplification = −0.5, low F_1 amplification = +0.5), variability (low variability = −0.5, high variability = +0.5), experiment (experiment 1 = −0.5, experiment 2 = +0.5), and all interactions; random slopes for target, variability, and experiment; and random intercepts by subject. The only fixed effect involving experiment that reached statistical significance was its interaction with variability, indicating that base rates of "eh" responses in experiment 1 were higher in the high variability in mean F_1 condition (54%) than in the low variability in mean F_1 condition (50%), but these were equal if not slightly higher in the low variability in mean F_3 condition of experiment 2 (54% versus 53%). The full model results are available in supplementary Table V.[1]

Responses in experiment 2 were also tested against those in experiment 2 of Assgari *et al.* (2019) to directly compare the perceptual consequences of variability in mean F_3 or mean f0 of context sentences, respectively. Responses were coded as before (/ɪ/ = 0, /ɛ/ = 1), and all main effects and interactions between target, filter, variability, and experiment were included. Filter and variability were sum-coded as above, and experiment was sum-coded as well (the present experiment = −0.5, 2019 experiment = +0.5). A model with random slopes for each fixed main effect and random intercepts for subjects did not converge, so random slopes were added iteratively and retained if the model converged and fit significantly improved. The final model had random slopes for target and random intercepts for subjects. The experiments exhibited distinct relationships between SCEs and variability (significant filter × variability × experiment interaction: $Z = -2.254$, $p = 0.024$), as variability in mean f0 modulated SCE magnitudes but variability in mean F_3 did not (cf. Table II). This finding is further supported by the filter × variability interaction being only marginally significant when averaging across these experiments ($Z = -1.719$, $p = 0.086$), as was observed in analyses following

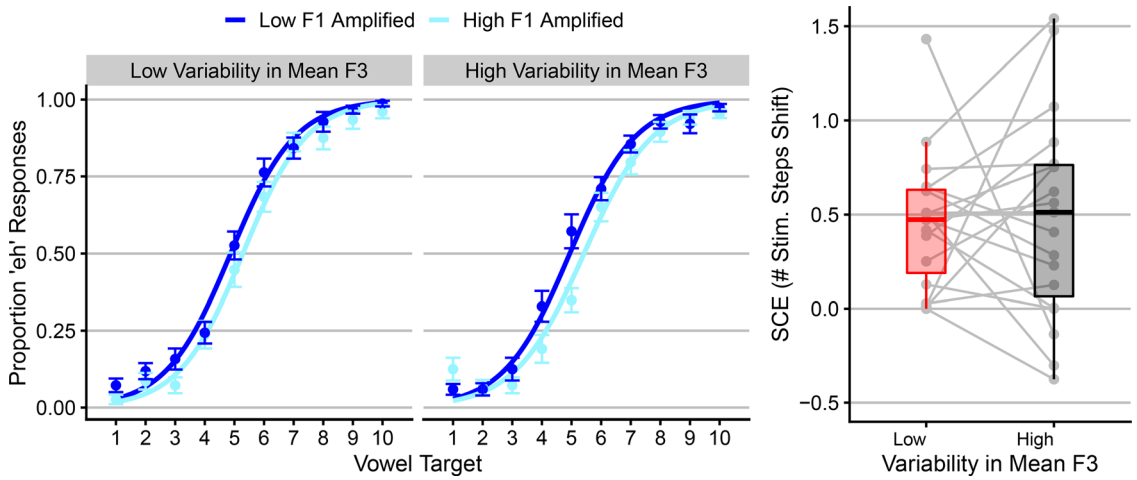

FIG. 3. (Color online) (Left) The mixed-effects model fit to listeners' responses as a function of vowel target in experiment 2. Dark blue lines represent responses following low-F_1-amplified context sentences; light blue lines represent responses following high-F_1-amplified context sentences. Circles depict mean proportions of "eh" responses; error bars depict one standard error of the mean. (Right) SCE magnitudes (the number of stimulus steps separating 50% points on the psychometric functions) calculated for each listener in each variability condition (red indicating low variability in mean F_3, black indicating high variability in mean F_3, consistent with coloration of Fig. 1). Each gray line connects one listener's SCEs in both variability conditions.

https://doi.org/10.1121/10.0011920

experiment 1. The full model results are available in supplementary Table VI.[1]

C. Discussion

Variability in the mean F_3 frequency of context sentences did not alter SCE magnitudes in vowel categorization. This was the same pattern of results observed in experiment 1. Statistical analyses indicated that neither SCEs nor their relationship with acoustic variability in the context sentences differed across experiments. Thus, neither the variability in mean F_1 frequencies of context sentences (experiment 1) nor in mean F_3 frequencies of context sentences (experiment 2) systematically altered SCE magnitudes in categorization of the target vowels /ɪ/-/ɛ/.

The magnitudes of SCEs were the primary metric to assess different perceptual impacts of variability in mean f0 versus variability in mean F_3 on vowel categorization. These different perceptual impacts were also evident in the slopes of the psychometric functions, as assessed by target × variability interactions in the mixed-effects models. This interaction was significant in Assgari *et al.* (2019), indicating psychometric function slopes were shallower in the high variability in mean f0 condition than the low variability in mean f0 condition. This interaction was not significant in experiment 2, indicating that variability in mean F_3 did not alter the slopes of the psychometric functions (Table II). In the mixed-effects analysis of these two experiments reported above, the target × variability × experiment interaction was statistically significant ($Z = -2.592$, $p = 0.010$), reinforcing the different patterns of psychometric slopes as functions of acoustic variability across experiments. These comparisons follow the same patterns as when experiment 1 was being compared to experiment 2 of Assgari *et al.* (2019). This offers yet further support to the notion that not all sources of acoustic variability in context sentences influence SCEs in vowel categorization equally.

The results of experiment 2 are also in contradiction with predictions made by the efficient coding hypothesis. Despite there being very different amounts of variability in mean F_3 frequencies across the two testing blocks, SCE magnitudes were unaffected. The broader implications of these findings are discussed in the Sec. IV.

IV. GENERAL DISCUSSION

When spectral properties differ across earlier (context) and later (target) sounds, speech categorization can become biased through SCEs. Previous studies have reported mixed results as to whether talker variability disrupts SCEs in vowel categorization. When the talkers who spoke context sentences were highly variable in their mean f0's, SCEs were diminished; when talkers varied in their mean F_1 frequencies, SCE magnitudes were unaffected (Assgari *et al.*, 2019). However, when testing effects of variability in mean F_1, Assgari *et al.* (2019) presented sentences that were originally selected based on measures of mean f0, underrepresenting the full range of F_1 variability across different talkers. Additionally, higher formants (such as F_3) may contribute differently to perception of speech sounds versus perception of talker characteristics as compared to lower formants (like F_1), so variability in these different spectral properties may have different impacts on (context effects in) speech perception. Here, context sentences were selected specifically based on mean F_1 measures (experiment 1) or mean F_3 measures (experiment 2) to more broadly explore perceptual sensitivity to acoustic variability in contexts when categorizing subsequent vowel sounds. Across experiments, SCEs occurred, but their magnitudes were unaffected by the amount of variability in mean F_1 or in mean F_3 in the context sentences. In conjunction with the findings of Assgari *et al.* (2019), context effects appear to be sensitive to acoustic variability in some aspects of talkers' voices (mean f0) but not others (mean F_1, mean F_3).

Recently heard sounds form a context for subsequent perception. In many studies of talker adaptation, the talker on a given trial forms a perceptual context for the stimulus heard on the next trial. When the talker repeats on the next trial (as in single-talker conditions), perception is facilitated (faster and/or more accurate response). When the talker changes on the next trial (as in mixed-talker conditions), perception is challenged (slower and/or less accurate response). From this perspective, studies of talker adaptation and acoustic context effects (such as SCEs) share certain similarities. Both measure perceptual sensitivity to context (within and/or across trials) and can be challenged by talker variability (especially regarding f0 characteristics as outlined in the Introduction). Talker variability has been shown to impair speech perception in a wide variety of tasks, including word identification (Mullennix *et al.*, 1989), word list recall (Goldinger *et al.*, 1991; Martin *et al.*, 1989), vowel monitoring (Barreda, 2012; Magnuson and Nusbaum, 2007), voice classification (Mullennix and Pisoni, 1990), lexical tone categorization (Zhang and Chen, 2016), and categorization of isolated phonemes (Assmann *et al.*, 1982). Results demonstrating that variability in the mean f0 of context sentences diminishes SCEs in vowel categorization (Assgari *et al.*, 2019) are consistent with this literature. However, one sizable difference between these tasks merits further discussion. In most studies measuring perceptual sensitivity to talker variability, target stimuli comprise all of the sounds listeners hear in the task. In this arrangement, talker variability directly impacts perception because the (target) stimuli are spoken by the same talker or different talkers. In acoustic context effect studies such as the present experiment [as well as Assgari and Stilp (2015) and Assgari *et al.* (2019)], context stimuli might be spoken by different talkers, but a single set of target stimuli spoken by one talker is presented. Put another way, talker variability exists in the context stimuli but not in the target stimuli. Categorization of the target sounds does not require any attention or response to the context stimuli, yet their acoustic characteristics (spectral properties that elicit SCEs and variability in mean f0 that diminishes the magnitudes of SCEs) are still perceptually influential. Thus, the impact of talker variability

on speech perception is broad, as variability entirely within the target stimuli (as in many talker adaptation paradigms) or beyond the target stimuli (in the context stimuli, as in the present experimental paradigm) can incur processing costs that challenge performance.

Both f0 and formant frequencies vary considerably from talker to talker (Hillenbrand *et al.*, 1995; Peterson and Barney, 1952) and play important roles in distinguishing talkers' voices and/or sexes, but their relative contributions are still debated. In studies that manipulated both f0 and formant frequencies, some have suggested that f0 is more influential for talker identification (Compton, 1963; Walden *et al.*, 1978; Hillenbrand and Clark, 2009; Baumann and Belin, 2010). For example, Compton (1963) reported that talker identification was substantially reduced when high-pass filtering vowels at a cutoff frequency of 1020 Hz, but low-pass filtering vowels at that same cutoff frequency had no effect on performance. Hillenbrand and Clark (2009) reported that shifting both f0 and formants up to higher frequencies was effective for changing perceived talker sex from male to female (and vice versa), but shifting only f0 was more effective than shifting only formants. However, other studies have suggested that formant frequencies are more influential than f0 for talker discrimination (Coleman, 1971; Childers and Wu, 1991; Bachorowski and Owren, 1999; Lavner *et al.*, 2000). For example, Childers and Wu (1991) reported slightly higher accuracies for distinguishing talker gender using formant frequencies compared to using f0. Bachorowski and Owren (1999) analyzed 2500 tokens of /ɛ/ from naturally produced speech by 125 talkers. Classification of talkers by discriminant analyses depended primarily on acoustic properties affiliated with vocal tract filtering (i.e., formant frequencies). Still other studies concluded that f0 and formant frequencies were roughly equally influential (LaRiviere, 1975; Smith and Patterson, 2005; Assmann *et al.*, 2006). Finally, Van Lancker *et al.* (1985) posited that no single signal property is expected to be critical for identifying all voices. Consensus about whether f0 or formant frequencies are more important (or equally important) in these tasks may be lacking. However, perception of talker identity and perception of talker sex are very different tasks; context effects in speech sound categorization are different further still. It is quite possible that whether f0 or formant frequencies are more perceptually influential is sensitive to the type(s) of variability present, the degree(s) of variability tested, the task, and perhaps also the stimuli; these considerations merit focused attention for research going forward.

Predictions drew on the efficient coding hypothesis (Attneave, 1954; Barlow, 1961), where structure in the sensory environment is predicted to facilitate perceptual performance, whereas the lack of structure is predicted to challenge performance. Specifically, SCE magnitudes were predicted to be larger in conditions where acoustic variability was low/where stimulus structure was higher (low variability in mean F_1, low variability in mean F_3) than in conditions where acoustic variability was high/where stimulus structure was lower (high variability in mean F_1, high variability in mean F_3). These predictions were motivated by previous findings that SCE magnitudes were larger when context sentences exhibited low variability in mean f0 compared to when they exhibited higher variability in mean f0 (Assgari and Stilp, 2015; Assgari *et al.*, 2019) as well as faster response times in speeded word identification as f0 variability decreased/stimulus structure increased (fastest to single-talker stimuli, slower to mixed talkers with low f0 variability, slowest to mixed talkers with high f0 variability; Stilp and Theodore, 2020). However, the present results do not conform to the predictions of efficient coding: variability in mean F_1 (experiment 1) or mean F_3 (experiment 2) differed across testing blocks, but variability in mean f0 (which does modulate SCE magnitudes) was matched across blocks in each experiment. If perceptually salient structure in the acoustic environment (context talkers' mean f0's) is held constant, perception can be expected to operate similarly in those cases. If structure exists but it is not relevant for the task at hand (context talkers' mean F_1s or mean F_3s), perception might not be expected to capitalize on it. Because experiments are often designed to measure the effects of relevant structure on perception and not irrelevant structure, it is possible that the efficient coding literature generally portrays the case that perception readily exploits all available stimulus structure. The present results question that notion, as structure was clearly present in the context sentences, but it did not affect subsequent vowel categorization. This pattern is analogous to word recognition and memory studies conducted by Pisoni and colleagues (Sommers *et al.*, 1994; Nygaard *et al.*, 1995; Bradlow *et al.*, 1999), in which variation in talkers' speaking rates impaired listeners' performance, but variation in amplitude did not. While they conclude that not all sources of variability in the speech signal reduce perceptual performance, Sommers *et al.* (1994) expressly note that "comparing the perceptual consequences of variations along different stimulus dimensions must be interpreted with caution" (p. 1321) owing to potential differences in perceptual saliency, ranges of variability tested, and other methodological concerns. Future research should more carefully distinguish between perceptually relevant structure (in a given task) from perceptually irrelevant structure to highlight when and how perception leverages this information to make processing efficient. Such investigations will clarify the utility of the efficient coding perspectives of speech perception going forward.

The question as to why this particular pattern of results was observed is an open one. The f0 of a talker's voice sets the resolution (spacing) of harmonics. While this has long been a consideration for the resolution with which formant peaks are specified (Fant, 1970), it also impacts the resolution of +5-dB spectral peaks added to the context sentences to produce SCEs. When mean f0 is relatively consistent from trial to trial, the harmonic resolution of the spectral peak in the low-F_1 (100–400 Hz) or high-F_1 (550–850 Hz) region is relatively consistent as well. When mean f0 is highly variable from trial to trial, the resolution of these

spectral peaks can vary considerably, from a more precise definition (multiple harmonics, as produced by a talker with a lower f0) to a much sparser definition (a single harmonic, as produced by a talker with a high f0). This variability in spectral peak resolution and/or the effectiveness of a sparsely defined spectral peak in producing an SCE might underlie the results of Assgari *et al.* (2019). Conversely, variation in mean F_1 or mean F_3 frequencies of context sentences would not affect the resolution of the spectral peaks added to low-F_1 or high-F_1 regions by filtering. Resolution in those regions is driven by f0, but in the present experiments, variability in mean f0 was matched across these conditions. Additionally, several reports indicate that variability in talkers' f0 characteristics challenges perception (Goldinger, 1996; Magnuson and Nusbaum, 2007; Stilp and Theodore, 2020), but comparable tests of the perceptual impact of variability in mean F_1 or mean F_3 in other tasks are lacking. Variability in the formant frequencies of target words is sufficient to slow recognition (when variability in f0 is matched across blocks), but these processing costs are smaller than when f0 variability is also present (Drown and Theodore, 2020). Targeted experimentation is needed to elucidate the mechanisms underlying these effects of f0 variability and the generalizability of this pattern of results (greater perceptual consequences for f0 variability than formant variability) to other tasks.

ACKNOWLEDGMENTS

This study was presented as the first author's Culminating Undergraduate Experience in the Department of Psychological and Brain Sciences at the University of Louisville. We wish to thank Emma Hatter, Lilah Kahloon, Samantha Schawe, and Chloe Sharpe for their assistance in stimulus analysis and preparation. All data and analysis scripts are available at https://osf.io/2nvur/.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0011920 for tables listing stimulus information and the full results of the mixed-effects models comparing responses during the present experiment to those in experiments 2 and 3 in Assgari *et al.* (2019).

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (**2020**). "Gorilla in our midst: An online behavioral experiment builder," Behav. Res. **52**(1), 388–407.
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.
Assgari, A. A., Theodore, R. M., and Stilp, C. E. (**2019**). "Variability in talkers' fundamental frequencies shapes context effects in speech perception," J. Acoust. Soc. Am. **145**(3), 1443–1454.
Assmann, P. F., Dembling, S., and Nearey, T. M. (**2006**). "Effects of frequency shifts on perceived naturalness and gender information in speech," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, September 17–21, Pittsburgh, PA, pp. 889–892.
Assmann, P. F., and Nearey, T. M. (**2008**). "Identification of frequency-shifted vowels," J. Acoust. Soc. Am. **124**(5), 3203–3212.
Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**(4), 975–989.
Attneave, F. (**1954**). "Some informational aspects of visual perception," Psychol. Rev. **61**(3), 183–193.
Bachorowski, J.-A., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **106**(2), 1054–1063.
Barlow, H. B. (**1961**). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, edited by W. A. Rosenblith (MIT, Cambridge, MA), pp. 53–85.
Barreda, S. (**2012**). "Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis," J. Acoust. Soc. Am. **132**(5), 3453–3464.
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," https://cran.r-project.org/web/packages/lme4/index.html (Last viewed June 21, 2022).
Baumann, O., and Belin, P. (**2010**). "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," Psychol. Res. **74**(1), 110–120.
Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer (version 6.1), [computer program]," http://www.praat.org (Last viewed July 13, 2019).
Bradlow, A. R., Nygaard, L. C., and Pisoni, D. B. (**1999**). "Effects of talker, rate, and amplitude variation on recognition memory for spoken words," Percept. Psychophys. **61**(2), 206–219.
Childers, D. G., and Wu, K. (**1991**). "Gender recognition from speech. Part II: Fine analysis," J. Acoust. Soc. Am. **90**(4), 1841–1856.
Choi, J. Y., Hu, E. R., and Perrachione, T. K. (**2018**). "Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing," Atten. Percept. Psychophys. **80**(3), 784–797.
Coleman, R. O. (**1971**). "Male and female voice quality and its relationship to vowel formant frequencies," J. Speech Hear. Res. **14**(3), 565–577.
Compton, A. J. (**1963**). "Effects of filtering and vocal duration upon the identification of speakers, aurally," J. Acoust. Soc. Am. **35**(11), 1748–1752.
Creelman, C. D. (**1957**). "Case of the unknown talker," J. Acoust. Soc. Am. **29**(5), 655.
Drown, L., and Theodore, R. M. (**2020**). "Effects of phonetic and indexical variability on talker normalization," J. Acoust. Soc. Am. **148**, 2504.
Fant, G. (**1970**). *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations* (Mouton de Gruyter, Berlin).
Field, D. J. (**1987**). "Relations between the statistics of natural images and the response properties of cortical cells," J. Opt. Soc. Am. A **4**(12), 2379–2394.
Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (**1990**). "DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," NIST Order No. PB91-505065, National Institute of Standards and Technology, Gaithersburg, MD.
Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (**2001**). "Edge co-occurrence in natural images predicts contour grouping performance," Vision Res. **41**(6), 711–724.
Gervain, J., and Geffen, M. N. (**2019**). "Efficient neural coding in auditory and speech perception," Trends Neurosci. **42**(1), 56–65.
Goldinger, S. D. (**1996**). "Words and voices: Episodic traces in spoken word identification and recognition memory," J. Exp. Psychol. Learn. Mem. Cogn. **22**(5), 1166–1183.
Goldinger, S. D., Pisoni, D. B., and Logan, J. S. (**1991**). "On the nature of talker variability effects on recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **17**(1), 152–162.
Hillenbrand, J. M., and Clark, M. J. (**2009**). "The role of f0 and formant frequencies in distinguishing the voices of men and women," Atten. Percept. Psychophys. **71**(5), 1150–1166.
Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111.
Johnson, K., and Sjerps, M. J. (**2021**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, 2nd edited by J. S. Pardo, L. C. Nygaard, R. E. Remez, and D. B. Pisoni (Wiley, New York), pp. 145–176.
Kluender, K. R., Stilp, C. E., and Kiefte, M. (**2013**). "Perception of vowel sounds within a biologically realistic model of efficient coding," in *Vowel Inherent Spectral Change*, edited by G. S. Morrison and P. F. Assmann (Springer, Berlin), pp. 117–151.

Kluender, K. R., Stilp, C. E., and Llanos, F. (**2019**). "Longstanding problems in speech perception dissolve within an information-theoretic perspective," Atten. Percept. Psychophys. **81**(4), 861–883.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**(1), 98–104.
Lammert, A. C., and Narayanan, S. S. (**2015**). "On short-time estimation of vocal tract length from formant frequencies," PLoS ONE **10**(7), e0132193.
LaRiviere, C. (**1975**). "Contributions of fundamental frequency and formant frequencies to speaker identification," Phonetica **31**(3), 185–197.
Lavner, Y., Gath, I., and Rosenhouse, J. (**2000**). "Effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," Speech Commun. **30**(1), 9–26.
Long, J. A. (**2019**). "Interactions: Comprehensive, user-friendly toolkit for probing interactions. R package version 1.1.3," https://cran.r-project.org/web/packages/interactions/index.html (Last viewed June 21, 2022).
Magnuson, J. S., and Nusbaum, H. C. (**2007**). "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," J. Exp. Psychol. Hum. Percept. Perform. **33**(2), 391–409.
Martin, C. S., Mullennix, J. W., Pisoni, D. B., and Summers, W. V. (**1989**). "Effects of talker variability on recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **15**(4), 676–684.
Mullennix, J. W., and Pisoni, D. B. (**1990**). "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**(4), 379–390.
Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.
Nordström, P. E., and Lindblom, B. (**1975**). "A normalization procedure for vowel formant data," in *Proceedings of the Eighth International Congress of Phonetic Sciences*, August 17–23, Leeds, UK.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1995**). "Effects of stimulus variability on perception and representation of spoken words in memory," Percept. Psychophys. **57**(7), 989–1001.
Olshausen, B. A., and Field, D. J. (**1996**). "Natural image statistics and efficient coding," Network **7**(2), 333–339.
Peterson, G. E. (**1951**). "The phonetic value of vowels," Language **27**(4), 541–553.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**(2), 175–184.
R Development Core Team (**2021**). "*R: A Language and Environment for Statistical Computing*," R Foundation for Statistical Computing, Vienna, Austria.
Ruderman, D. L., Cronin, T. W., and Chiao, C. C. (**1998**). "Statistics of cone responses to natural images: Implications for visual coding," J. Opt. Soc. Am. A **15**(8), 2036–2045.
Schwartz, O., and Simoncelli, E. P. (**2001**). "Natural signal statistics and sensory gain control," Nat. Neurosci. **4**(8), 819–825.
Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**(5), 3177–3186.
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.
Stilp, C. E. (**2020**). "Acoustic context effects in speech perception," Wiley Interdiscip. Rev. Cogn. Sci. **11**(1), 1–18.
Stilp, C. E., and Theodore, R. M. (**2020**). "Talker normalization is mediated by structured indexical information," Atten. Percept. Psychophys. **82**, 2237–2243.
Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (**1994**). "Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude," J. Acoust. Soc. Am. **96**(3), 1314–1324.
Tkačik, G., Prentice, J. S., Victor, J. D., and Balasubramanian, V. (**2010**). "Local statistics in natural scenes predict the saliency of synthetic textures," Proc. Natl. Acad. Sci. U.S.A. **107**(42), 18149–18154.
Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985**). "Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices," J. Phon. **13**(1), 19–38.
Wakita, H. (**1977**). "Normalization of vowels by vocal-tract length and its application to vowel identification," IEEE Trans. Acoust. Speech Signal Process. **25**, 183–192.
Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., and Schwartz, D. M. (**1978**). "Correlates of psychological dimensions in talker similarity," J. Speech Hear. Res. **21**(2), 265–275.
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.
Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (**2017**). "Headphone screening to facilitate web-based auditory experiments," Atten. Percept. Psychophys. **79**(7), 2064–2072.
Zhang, C., and Chen, S. (**2016**). "Toward an integrative model of talker normalization," J. Exp. Psychol. Hum. Percept. Perform. **42**(8), 1252–1268.